11-21-2007

A BAYESIAN MODEL FOR CROSS-STUDY DIFFERENTIAL GENE EXPRESSION

Robert B. Scharpf
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, rscharpf@jhsph.edu

Hakon Tjelemeland
Department of Mathematical Sciences, Norwegian University of Science and Technology

Giovanni Parmigiani
The Sydney Kimmel Comprehensive Cancer Center, Johns Hopkins University & Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Andrew B. Nobel
Department of Statistics, University of North Carolina, Chapel Hill

A Bayesian model for cross-study differential gene expression

Robert B. Scharpf[1], Håkon Tjelmeland[2], Giovanni Parmigiani[1,3], and Andrew B. Nobel[4]

[1] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205

[2] Department of Mathematical Sciences, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway

[3] The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD 21205

[4] Department of Statistics, University of North Carolina, Chapel Hill, NC 27599.

Abstract

In this paper we define a hierarchical Bayesian model for microarray expression data collected from several studies and use it to identify genes that show differential expression between two conditions. Key features include shrinkage across both genes and studies; flexible modeling that allows for interactions between platforms and the estimated effect, and for both concordant and discordant differential expression across studies. We evaluated the performance of our model in a comprehensive fashion, using both artificial data, and a "split-sample" validation approach that provides an agnostic assessment of the model's behavior not only under the null hypothesis but also under a realistic alternative. The simulation results from the artificial data demonstrate the advantages of a Bayesian model. Compared to a more direct combination of t- or SAM-statistics, the $1-\text{AUC}$ values for the Bayesian model is roughly half of the corresponding values for the t- and SAM-statistics. Furthermore, the simulations provide guidelines for when the Bayesian model is most likely to be useful. Most noticeably, in small studies the Bayesian model generally outperforms other methods when evaluated by AUC, FDR, and MDR across a range of simulation parameters, and this difference diminishes for larger sample sizes in the individual studies. The split-study validation illustrates appropriate shrinkage of the Bayesian model in the absence of platform-, sample-, and annotation-differences that otherwise complicate experimental data analyses. Finally, we fit our model to four breast cancer studies employing different technologies (cDNA and Affymetrix) to estimate differential expression in estrogen receptor positive tumors versus negative ones. Software and data for reproducing our analysis are publicly available.

1 Introduction

Microarray technologies that simultaneously measure transcriptional activity in a very large number of genes have been widely used in biology and medicine in the last decade, and the resulting data is often publicly available. To increase the reliability and efficiency of biological investigations, it can be critical to combine data from several studies. However, when considering multiple studies, variation in the measured gene expression levels is caused not only by the biological differences of interest and natural variation in gene expression within a phenotype, but also by technological and laboratory-based differences between studies (Irizarry et al., 2005; Consortium et al., 2006; Kerr, 2007). Two of the most important difficulties are the presence of both absolute and relative expression measurements depending on the technology, and the challenges associated with cross referencing measurements made by different technologies to the genome and to each other (Zhong et al., 2007). Despite these difficulties, the results of combined analysis clearly demonstrate the potential for increased statistical power and novel discovery by combining data from several studies (Wu et al., 2002; Rhodes et al., 2002; Tomlins et al., 2005).

Most statistical work to date on combining microarray studies focused on identifying genes that exhibit differential expression across two experimental conditions or phenotypes. We consider this problem here as well. There is now a substantial literature on Bayesian approaches to assessing differential expression across two or more experimental conditions within a single study. Both empirical and fully Bayesian models have been proposed, including parametric (Baldi and Long, 2001; Newton et al., 2001; Lönnstedt and Speed, 2002; Pan, 2002; Gottardo et al., 2003; Ishwaran and Rao, 2003; Kendziorski et al., 2003; Ishwaran and Rao, 2005; Tseng et al., 2001; Bröet et al., 2002; Ibrahim et al., 2002; Townsend and Hartl, 2002), semi-parametric (Newton et al., 2004) and non-parametric (Efron et al., 2001; Do et al., 2005) models. In each of these papers, a critical issue is shrinkage, and in particular borrowing strength across genes when estimating the gene-specific variance across samples. It is well established that shrinkage of the variance estimates provides worthwhile enhancements to single study analysis of differential expression (Dongmei Liu and Caffo, 2004).

There are several natural approaches for combining information from multiple microarray studies. One is to compute, separately for each study, statistics that summarize the relationship between each gene and the phenotype of interest. These may then be combined using methodologies such as

those originally devised to integrate published results in meta-analysis (Hedges and Olkin, 1985). While initial efforts in this direction have considered combination of p-values (Rhodes et al., 2002), subsequent papers have focused on the more efficient strategy of combining effect sizes (Ghosh et al., 2003; Wang et al., 2004; Garrett-Mayer et al., 2007). At the opposite extreme of study combination are cross-study normalization methods (Wu et al., 2002; Parmigiani, 2002; Shen et al., 2004; Rhodes et al., 2004; Hayes et al., 2006; Johnson et al., 2007; Choi et al., 2007) that consider directly the sample-level measurements within each study, and merge these into a single data set, to which standard single-study analysis can be applied. A third approach, intermediate between the two above, is to integrate information about differential expression from the available studies using a joint stochastic model for all the available data (Choi et al., 2003; Conlon et al., 2006; Jung et al., 2006; Conlon, 2007; Conlon et al., 2007), in which only selected features of each study, such as parameters that capture the relationship between genes and phenotypes, are assumed to be related across studies. This perspective has the potential to offer additional efficiency over integration of summary statistics, and to allow for a more comprehensive treatment of uncertainty. At the same time it models the cross-study integration in a way that is tailored to the problem of interest, and potentially relies on fewer assumptions than direct data integration.

In this article we adopt this latter, intermediate approach, and propose a fully Bayesian hierarchical model to identify genes that exhibit differential expression between two experimental conditions, and across multiple studies. In this context, use of a fully Bayesian model has several desirable features. The model borrows strength across both genes and studies and can thereby provide better estimates of the gene-specific means, variances and effects. The model yields, through simulation, posterior probability distributions for all unobserved quantities. These distributions can be used to quantify the uncertainty of any parameter in the model, or to make joint inferences about multiple genes. Lastly, for each gene, the model yields the posterior probability that the gene is differentially expressed.

While the work of Conlon and colleagues considers several of these issues, its primary strength is in the combination of multiple studies from the same technology. We expand this and related work to address multi-platform analysis via several technical generalizations that are described in detail in the methods and reviewed in the discussion. These include: modeling of overall cross-platform correlations to allow shrinkage to be stronger across pairs of studies that are generally more

concordant; modeling of both the mean of a gene and its phenotypic interactions with sufficient flexibility to avoid distributional modeling of the main effects; allowing for interactions between the technology and the effect; modeling of an adaptive smooth dependence between effects and the variance terms.

Perhaps the most radical difference between our approach and all its predecessor is the attention given to discordant differential expression. This occurs when a gene is more highly expressed in one phenotype than the other in some studies, while the opposite is observed in other studies. Earlier approaches would discount the gene: the high cross-study variance and cancelation of overall effects would likely position it with the uninteresting genes. However, across many meta-analyses, we have observed an excess of these discordant genes compared to what would have been predicted by chance alone, as captured by permutation of phenotype labels. When implementing shrinkage strategies, reliable assessments of concordant differential expression, which is typically of primary interest, must therefore account for the possibility of discordant differential expression across studies. We implement this by introducing a gene-specific indicator of whether a gene is different across conditions in all three studies, but then we allow these differences to be gene and study specific.

While concordant differentially expressed genes remain the primary focus of the analysis, discordant genes can reveal important biological or technological information, and it is useful to identify them and report them. This is for at least two reasons: first, given the heterogeneous experimental designs that are encountered in microarrays, a discordant effect for a set of important genes may be the result of genetic heterogeneity of the samples across study. For a simple example, consider the comparison of administering or not administering a certain drug in two studies which, unbeknownst to the investigators, use strains of animals where the sets of biochemical pathways activated by the drug are not the same. Then certain genes' expression may be increased by the drug in one strain and decreased in the other. Another reason discordance is important is that the cross-referencing of genes across studies is typically gene-centric. However, as many as 40 to 60% of genes are able to produce multiple alternative transcripts (Modrek and Lee, 2002), whose expression may be positively or, as is common, negatively correlated. For example one transcript may be made primarily under normal conditions, while the other may be made mainly in response to stress. When two technologies measure a gene's expression by targeting portions of that gene that are associated

with different transcripts that are negatively correlated with each other, discordant effects will be observed. In either case, important insight about technology, study designs and potentially the genetics of alternative splicing can potentially be gained by following up on discordant genes.

The paper is organized as follows. In Section 2 we describe our Bayesian model and in Section 3 we define a Markov chain Monte Carlo algorithm for simulation from the resulting posterior distribution. Section 4 describes statistics from the Bayesian model that can be used to quantify differential expression, as well as alternative approaches for quantifying differential expression in the context of multiple studies. The datasets used in the simulation and experimental data example are descibed in Section 5. Sections 6 and 7 present results when applying our model to simulated and real data with comparisons to alternative methods. Concluding remarks are in Section 8. The software for fitting our Bayesian model is freely available from Bioconductor.

2 Bayesian hierarchical cross-study model

In this section we introduce some basic notation and our Bayesian model. The resulting method for cross-study assessment of differential expression will be referred to as XDE.

2.1 Notation and basic assumptions

In what follows we use p and q to index studies (i.e. data sets), g to index genes, and s to index samples (arrays) within each study. Let x_{gsp} denote the observed expression value for gene g and sample s in study p. Let P denote the number of available studies, G the number of common genes and S_p the number of samples in study p. Thus the observed expression values are

$$\{x_{gsp} : g = 1, \ldots, G;\ s = 1, \ldots, S_p;\ p = 1, \ldots, P\}.$$

We assume that each study has been suitably normalized (and if necessary log-transformed) so that the mean expression value for each study is zero and the expression values for a given gene are approximately Gaussian under each condition. We restrict our analysis to the set of common genes in the available studies, though our model formulation can easily be extended to a situation in which there is substantial overlap, but not complete agreement, between the gene sets in different

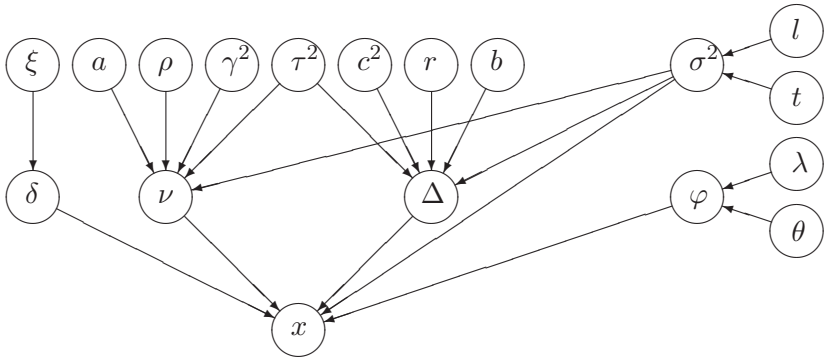


Figure 1: A graphical model representation of the hierarchical Bayesian model defined for the microarray data sets.

studies.

In the model described below, each sample is assumed to belong to one of two possible conditions or phenotypes. Let $\psi_{sp} \in \{0, 1\}$ denote the phenotype of sample s in study p. (An example in which ψ_{sp} represents the estrogen receptor of breast tumors is presented in Section 7.) In order to model differential expression, we assume that, for a subset of the available genes, the expression value x_{gsp} has a different mean value in samples where $\psi_{sp} = 0$ than in samples where $\psi_{sp} = 1$. Furthermore, for each gene we allow only two possible states of differential expression, indexed by a binary parameter δ_g: either the expression values of gene g are differentially expressed in all P studies ($\delta_g = 1$), or they are not differentially expressed in any of the studies ($\delta_g = 0$).

2.2 Bayesian model

We define a hierarchical Bayesian model for the expression values x_{gsp}. In the following discussion the graphical model representation in Figure 1 can be used as a reference.

At the lowest level we assume the expression values x_{gsp}, conditional on some unobserved parameters, are independent and have a Gaussian distribution. For genes that are not differentially expressed, ν_{gp} denotes the mean value of x_{gsp}, i.e., the mean value may be different for different genes and studies, but is the same mean for all samples in the same study. By contrast, differentially expressed genes have different means under the two phenotypic conditions. When $\delta_g = 1$ the mean of gene g in study p is equal to $\nu_{gp} - \Delta_{gp}$ and $\nu_{gp} + \Delta_{gp}$ for samples with $\psi_{sp} = 0$ and $\psi_{sp} = 1$, respectively. Thus Δ_{gp} represents half the average difference between expression levels across phenotypes for gene g in study p. By allowing Δ to depend on both g and p we acknowledge that

the measured magnitude of an effect may depend on the technology. We impose no restriction that the Δ_{gp}s should have the same sign across studies, thus allowing for the possibility of discordant differential expression. We also allow the variance of x_{gsp} to depend on the gene g, the study p, and the phenotypic condition ψ_{sp}. Let σ^2_{g0p} and σ^2_{g1p} denote the variances of x_{gsp} for samples with $\psi_{sp} = 0$ and $\psi_{sp} = 1$, respectively. Our basic model may be written as follows:

$$x_{gsp} \mid \nu_{gp}, \delta_g, \Delta_{gp}, \sigma^2_{g0p}, \sigma^2_{g1p} \;\sim\; \mathrm{N}\left(\nu_{gp} + \delta_g(2\psi_{sp} - 1)\Delta_{gp}, \sigma^2_{g\psi_{sp}p}\right). \tag{1}$$

At the next level in the model specification, prior distributions are selected for the model parameters. We begin by discussing the priors for

$$\boldsymbol{\nu}_g = (\nu_{g1}, \ldots, \nu_{gP})^T \quad \text{and} \quad \boldsymbol{\Delta}_g = (\Delta_{g1}, \ldots, \Delta_{gP})^T$$

which represent, respectively, the means and offsets for gene g. The vectors $\boldsymbol{\nu}_g$ and $\boldsymbol{\Delta}_g$ are assumed to be independent across genes, and independent of each other. Furthermore, for every gene g, it is assumed that $\boldsymbol{\nu}_g$ and $\boldsymbol{\Delta}_g$ have a multiGaussian distribution. As we have required that the expression values of each gene be centered around zero, we set the mean of $\boldsymbol{\nu}_g$ and $\boldsymbol{\Delta}_g$ equal to zero as well. Let Σ_g and R_g denote the covariance matrices of $\boldsymbol{\nu}_g$ and $\boldsymbol{\Delta}_g$, respectively, so that

$$\boldsymbol{\nu}_g \sim \mathrm{N}(0, \Sigma_g) \quad \text{and} \quad \boldsymbol{\Delta}_g \sim \mathrm{N}(0, R_g). \tag{2}$$

To specify the covariance matrices of Σ_g and R_g we adopt the strategy advocated in Barnard et al. (2000), namely, to assign independent prior distributions to the standard deviation and the correlation matrix of each quantity (see below for more details).

When modeling normal location and scale parameters in a hierarchical way, as we do here, two modeling choices are common. One is independence between scale and location, and the other is conjugacy. The latter is computationally more convenient, as it allows analytical expressions for the full conditional distribution. However, which of these two specifications fits better can vary from experiment to experiment, and in microarray analysis the fit is sensitive to the technology and normalization method used, see Liu et al. (2004). Recently, Caffo et al. (2004) proposed a more general family of models, one that encompasses both independence and conjugacy, by including a

single parameter that indexes the distribution of location given scale. Here we extend this idea in a natural way by introducing separate parameters for each individual study, and setting a common relative scale for Σ_g and R_g in each study. More specifically, the diagonal elements of Σ_g and R_g are given as follows:

$$(\Sigma_g)_{pp} = \gamma^2 \tau_p^2 \sigma_{gp}^{2a_p} \quad \text{and} \quad (R_g)_{pp} = c^2 \tau_p^2 \sigma_{gp}^{2b_g} \quad p = 1, \ldots, P \ . \tag{3}$$

Here $\sigma_{gp}^2 = \sqrt{\sigma_{g0p}^2 \, \sigma_{g1p}^2}$, the parameters $a_p, b_p \in [0, 1]$, and the parameters $\tau_p^2 > 0$ are such that $\tau_1^2 \cdot \ldots \cdot \tau_P^2 = 1$. Thus γ^2 and a_p control the overall scale and conjugacy of $\boldsymbol{\nu}_g$, respectively, while c^2 and b_p play analogous roles for $\boldsymbol{\Delta}_g$, and $\tau_1^2, \ldots, \tau_P^2$ control the relative scales of the different studies.

The correlation structure of Σ_g (and R_g) is assumed to be the same for all genes g. Let $[\rho_{pq}]_{p,q=1}^P$ and $[r_{pq}]_{p,q=1}^P$ denote the correlation matrices corresponding to Σ_g and R_g, respectively. Following Barnard et al. (2000), the prior distribution for $[\rho_{pq}]$ is obtained by beginning with a covariance matrix having an inverse Wishart distribution with ν_ρ degrees of freedom, and then integrating out its component variances. The prior distribution for $[r_{pq}]$ is of the same form, with ν_r degrees of freedom, and independent of the prior for $[\rho_{pq}]$.

At the next level in the hierarchical model specification, priors are placed on the hyper-parameters γ^2, c^2, τ_p^2, a_p and b_p. To enforce model parsimony, the prior distributions for a_p and b_p place positive probability mass at the values 0 and 1, corresponding to independence and conjugacy between location and scale, respectively. More specifically, independently for each study p, we let

$$\mathrm{P}(a_p = 0) = p_a^0 \ , \ \ \mathrm{P}(a_p = 1) = p_a^1 \ , \ \ a_p | a_p \in (0, 1) \sim \mathrm{Beta}(\alpha_a, \beta_a) \tag{4}$$

and

$$\mathrm{P}(b_p = 0) = p_b^0 \ , \ \ \mathrm{P}(b_p = 1) = p_b^1 \ , \ \ b_p | b_p \in (0, 1) \sim \mathrm{Beta}(\alpha_b, \beta_b). \tag{5}$$

Independent vague priors are assigned to the remaining hyper-parameters. For γ^2 we use an (improper) uniform distribution on $(0, \infty)$, and for c^2 a uniform distribution on $(0, c_{\max}^2)$. Note that an improper prior can not be used for c^2 as this may result in an improper posterior distribution. For $\tau_1^2, \ldots, \tau_P^2$ we assign a joint (improper) uniform distribution under the natural restrictions

$\tau_p^2 > 0, p = 1, \ldots, P$ and $\prod_{p=1}^{P} \tau_p^2 = 1$.

In order to have a fully defined Bayesian model, it remains to specify prior distributions for the differential expression indicators δ_g, and for the variances $\sigma^2_{g\psi p}$ used to define σ^2_{gp}. The indicators δ_g are assumed to be *a priori* independent, given a hyper-parameter ξ, with

$$\mathrm{P}(\delta_g = 1) = \xi \quad \text{and} \quad \xi \sim \mathrm{Beta}(\alpha_\xi, \beta_\xi). \tag{6}$$

The variances $\sigma^2_{g\psi p}$ are assumed to be independent for different genes g and studies p, given the other hyperparameters. However, σ^2_{g0p} and σ^2_{g1p} should be correlated for the same gene g and study p. To obtain this, we set

$$\sigma^2_{g0p} = \sigma^2_{gp}\varphi_{gp} \quad \text{and} \quad \sigma^2_{g1p} = \frac{\sigma^2_{gp}}{\varphi_{gp}}, \tag{7}$$

where σ^2_{gp} and φ_{gp} have independent gamma prior distributions with $\mathrm{E}[\sigma^2_{gp}] = l_p$, $\mathrm{Var}[\sigma^2_{gp}] = t_p$, $\mathrm{E}[\varphi_{gp}] = \lambda_p$ and $\mathrm{Var}[\varphi_{gp}] = \theta_p$. At the next level we assign independent (improper) uniform distributions on $(0, \infty)$ for each of the hyper-parameters $l_p, t_p, \lambda_p, \theta_p$, independently for $p = 1, \ldots, P$.

The above prescriptions fully define the hierarchical Bayesian model visualized in Figure 1. The observed quantities are the expression values x_{gsp} and the conditions ψ_{sp}. Conditioning on the observed values we get a posterior distribution for the unobserved parameters ξ, δ_g, a_p, ρ_{pq}, γ, τ_p^2, ν_{gp}, c^2, r_{pq}, b_p, Δ_{gp}, σ^2_{gp}, φ_{pg}, l_p, t_p, λ_p and θ_p. Hyper-parameters that have to be specified by the user are α_a, β_a, α_b, β_b, p^0_a, p^1_a, p^0_b, p^1_b, ν_ρ, ν_r, α_ξ, β_ξ and $c^2_{\max}$. Default hyperparameters provided in the R package *XDE* work well in most instances (see Table 4).

3 Posterior simulation

In order to evaluate the properties of the resulting posterior distribution, we adopt the Metropolis–Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) to generate realizations from it. Nice introductions to the Metropolis–Hastings algorithm can be found in Smith and Roberts (1993) and Dellaportas and Roberts (2003). The algorithm is iterative, and each iteration consists of two parts. First, potential new values for one or a number of parameters are proposed according to a proposal distribution. Second, the proposed values are accepted with a specified probability. Depending on the mathematical form of the distribution of interest, different proposal mechanisms can be

employed. In our posterior distribution we have seventeen types of parameters and to update these we combine seven types of proposal mechanisms. In the following we specify each of the proposal strategies used. In the description we use tilde to denote potential new values, e.g. δ_g and $\widetilde{\delta}_g$ are the current and potential new values of the differential expression indicator for gene g, respectively. Note that we restrict attention to the proposal distributions, as the acceptance probabilities are then uniquely defined by the Metropolis–Hastings setup. As with many Metropolis–Hastings proposal strategies, several of our proposal distributions include a "tuning" parameter that measures the amount of change proposed. In the following we consistently use the same symbol ε to denote all the tuning parameters, but, as the values used in our examples in Sections 6 and 7 suggest, one can of course use different values when updating the different parameters.

1. The full conditionals for $\boldsymbol{\nu}_g$, $\boldsymbol{\Delta}_g$, γ^2 and ξ have standard forms and we therefore use Gibbs steps (see the references given above) to update each of these separately. The full conditionals for $\boldsymbol{\nu}_g$ and $\boldsymbol{\Delta}_g$ both are multiGaussians, the full conditional for γ^2 is an inverse gamma distribution, and for ξ it is a beta distribution.

2. Separately, for each of the parameters a_p and b_p, we use a "truncated" random walk proposal. In particular, for a_p we do the following: if $a_p = 0$ we draw $\widetilde{a}_p$ from a uniform distribution on $[0, \varepsilon]$; if $a_p = 1$ we draw $\widetilde{a}_p$ uniformly on $[1 - \varepsilon, 1]$; and if $a_p \in (0, 1)$ we draw U from a uniform distribution on $[a_p - \varepsilon, a_p + \varepsilon]$ and set $\widetilde{a}_p = \min(1, \max(0, U))$. We note that this is a reversible jump type of proposal, and to get the correct acceptance probability, one needs to use the theory introduced in Green (1995).

3. Separately, for each of the parameters σ^2_{gp}, φ_{gp}, l_p, t_p, λ_p and θ_p, we propose a multiplicative change. In particular, for σ^2_{pq} we set $\widetilde{\sigma}^2_{pq} = u\sigma^2_{pq}$, where u is sampled from a uniform distribution on the interval $[1/(1+\varepsilon), 1+\varepsilon]$.

4. When updating $(\tau^2_1, \ldots, \tau^2_P)$ we must ensure that the product of the proposed new values equals unity. We do this by randomly selecting two of the components, p and q say, drawing U from a uniform distribution on $[1/(1+\varepsilon), 1+\varepsilon]$, and setting $\widetilde{\tau}^2_p = U\tau^2_p$ and $\widetilde{\tau}^2_q = \tau^2_q/U$.

5. A block Gibbs update is used for c^2 and all the $\boldsymbol{\Delta}_g$'s for genes that have $\delta_g = 0$.

6. Separately for each $g = 1, \ldots, G$, a block update is used for δ_g and $\mathbf{\Delta}_g$. First, the potential new value for δ_g is set by inverting the current value, i.e. $\tilde{\delta}_g = 1 - \delta_g$. Second, a potential new value for $\mathbf{\Delta}_g$ is sampled from the associated full conditional (given the potential new value $\tilde{\delta}_g$). The proposed values are then accepted or rejected jointly.

7. A block update is used for $[\rho_{pq}]$ and γ^2. A similar block update is used for $[r_{pq}]$ and c^2. For $[\rho_{pq}]$ and γ^2, potential new values for $[\rho_{pq}]$ are obtained via the transformation

$$\tilde{\rho}_{pq} = (1 - \varepsilon)\rho_{pq} + \varepsilon T_{pq}.$$

Here $[T_{pq}]$ is a correlation matrix which with probability one half is generated from the prior for $[\rho_{pq}]$, and with probability one half is set equal to unity on the diagonal with constant off diagonal elements. In the latter case, the value of the off diagonal elements is sampled from a uniform distribution on $(-1/(P-1), 1)$. Thereafter, the potential new value for γ^2 is sampled from the associated full conditional (given the potential new values $[\tilde{\rho}_{pq}]$). The proposed values are then accepted or rejected jointly.

4 Estimation of differential expression

In assessing the differential expression of genes across multiple studies, one naturally encounters a difficulty that is not present in single study analyses. This difficulty arises from the fact that a single differentially expressed gene g may be up-regulated in one or more studies, and down-regulated in others. When this occurs, we say that g is discordantly differentially expressed. If g is up-regulated in every study, or down-regulated in every study, we say that g is concordantly differentially expressed. Although concordant differential expression is the norm, discordance can arise from biological differences in the sample populations of each study, or from technological effects related to the design and implementation of specific array technologies. Discordance appears to be an unavoidable (and inconvenient) feature of multi-study analyses, one that comprehensive multi-study analyses should take into account.

4.1 Bayesian estimation

In our Bayesian model we have assumed that a gene is either differentially expressed in all of the studies or in none of the studies. Thus the indicator δ_g is summarizing information across studies. The basis for our cross-platform analysis of differential expression is the posterior mean of δ_g, equivalently the posterior probability that gene g is differentially expressed. This posterior mean is not analytically available, so in practice we have to generate samples from the posterior distribution, as discussed in Section 3, and estimate the posterior mean by the empirical mean of the simulated δ_g's.

Let $\mathrm{PM}_{\mathcal{E}}(g)$ denote the posterior mean of δ_g. We view $\mathrm{PM}_{\mathcal{E}}(g)$ as a measure of the evidence for the overall differential expression of g. In particular, one may classify a gene g as differentially expressed whenever $\mathrm{PM}_{\mathcal{E}}(g) > a$ for some threshold $a > 0$. Concordant and discordant differential expression can also be addressed in a direct way in the context of the Bayesian model described above. A gene g for which $\delta_g = 1$ is concordantly differentially expressed if each of its offsets Δ_{gp}, $p = 1, \ldots, P$ has the same sign, and is discordant if its offsets include both positive and negative values. Thus, indicators for concordant and discordant differential expression can be defined by

$$\mathcal{C}_g = \begin{cases} 1 & \text{if } \delta_g = 1 \text{ and all } \Delta_{gp}, p = 1, \ldots, P \text{ have the same sign,} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

and

$$\mathcal{D}_g = \begin{cases} 1 & \text{if } \delta_g = 1 \text{ and the } \Delta_{gp}, p = 1, \ldots, P \text{ do not all have the same sign,} \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

respectively. The posterior mean of each indicator can again be estimated by the empirical mean of the corresponding simulated quantities. Let $\mathrm{PM}_{\mathcal{C}}(g)$ and $\mathrm{PM}_{\mathcal{D}}(g)$ denote the corresponding posterior mean values. Then a gene g may be classified as concordantly or discordantly differentially expressed whenever $\mathrm{PM}_{\mathcal{C}}(g) > a$ or $\mathrm{PM}_{\mathcal{D}}(g) > a$ for some threshold $a > 0$.

4.2 Alternative methods

We consider three alternatives to *XDE* for estimating differential expression: the implementation of the Choi et al. (2003) random effects model in the R package *GeneMeta* (Gentleman et al., 2005),

and cross-study summaries of t- and SAM-statistics. While a comparison with the Conlon et al. (2006) paper would be interesting, software for fitting this model to expression data is not readily available.

The study-specific statistics from which we derive cross-study summaries of differential expression are the Welch t-statistic (t_{gp}), $\mathtt{SAM}_{gp}$ (Tusher et al., 2001), and a standardized unbiased estimate for effect size, $\mathtt{z}_{gp}$ (Hedges and Olkin, 1985), discussed by Choi et al. (2003) in the context of a cross-study microarray analysis. The Welch t-statistic allows for unequal variances between the phenotypes, whereas the z statistic uses a pooled estimate that assumes equal variance between the phenotypes. In contrast to the t- and z-statistics, the SAM statistic downweights genes with small variance, favoring genes with larger effect sizes. We hereafter generically denote the study-specific statistics by $\mathbf{U}_g = (U_{g1}, \ldots, U_{gP})$, and cross-study summaries of differential expression by non-negative statistics $u_*(g)$, where the subscript indicates whether the statistic measures overall differential expression ($\mathcal{E}$), concordant differential expression ($\mathcal{C}$), or discordant differential expression ($\mathcal{D}$). A gene g may then be classified as being appropriately differentially expressed if the corresponding statistic $u_*(g)$ exceeds a fixed constant $a > 0$.

For evaluating overall differential expression, we follow the discussion of Garrett-Mayer et al. (2004), and combine the elements of $\mathbf{U}_g$ in a linear fashion to obtain a statistic suitable for assessing differential expression:

$$\mathrm{u}_{\mathcal{E}}(g) \equiv \alpha_1|U_{g1}| + \ldots + \alpha_P|U_{gP}|, \text{ where} \tag{10}$$
$$\alpha_p \equiv \frac{\mathrm{L}_p\sqrt{S_p}}{\sum_{i=1}^{P}\mathrm{L}_i\sqrt{S_i}} \text{ for } p \in \{1, \ldots, P\},$$

Here L is the covariance loading from the first principal components and S is the number of samples. Summary measures of concordance for t_{gp} and SAM were obtained by

$$\mathrm{u}_{\mathcal{C}}(g) \equiv |\alpha_1 U_{g1} + \ldots + \alpha_P U_{gP}|.$$

As an alternative, we also used the combined (across studies) estimate of effect size from the random effects model proposed by Choi et al. (2003) directly. We denote this statistic by $\mathrm{z}_{\mathcal{C}}(g)$. The 'borrowing of strength' in the estimation of $\mathrm{z}_{\mathcal{C}}(g)$ is strictly across studies (as opposed to across

genes and studies), as the study-specific effect sizes for a given gene are assumed to be a draw from a Gaussian distribution in the second level of the random effects model. Assessments of discordant differential expression using the t_{gp}, `SAM` and z statistics are obtained by

$$\mathrm{u}_{\mathcal{D}}(g) \equiv \begin{cases} \mathrm{u}_{\varepsilon}(g) & \mathrm{sign}(U_{g1}) = \cdots = \mathrm{sign}(U_{gP}) \\ -1 \times \mathrm{u}_{\varepsilon}(g) & \text{otherwise.} \end{cases}$$

5 Datasets and Software

The package *XDE* contains the software used to fit the Bayesian hierarchical model, as well as convenient methods to compute the alternative statistics described in this paper. To facilitate reproducibility of the analyses in Sections 6 and 7, the datasets used in this manuscript are available as R objects in the R packages indicated below.

Lung cancer datasets. Parmigiani et al. (2004) applied a robust multichip average (rma; Irizarry et al. (2003)) separately to the 203 samples in `Harvard` (Bhattacharjee et al., 2001) (12,453 probesets on the Affymetrix Hu95a platform) and 108 samples in `Michigan` (Beer et al., 2002) (6663 probesets on the Affymetrix HG6800 platform). Intensity ratios from the Cy5 and Cy3 channels for the 68 samples in the `Stanford` (Garber et al., 2001) dataset (23,100 features on the cDNA platform) were log-transformed. Following normalization, probesets (Affymetrix) and image clone identifiers (cDNA) in each platform were mapped to UniGene identifiers. Many-to-one mappings (multiple probes map to one UniGene identifier) were averaged and one-to-many mappings were excluded. The studies were then merged by UniGene identifiers, resulting in a common set of 3,171 features. The three lung cancer datasets are available in the R package *lungExpression* available on the Bioconductor website (`http://www.bioconductor.org`).

Breast cancer datasets. Four breast cancer studies containing phenotypic data on estrogen receptor (ER) status (Sorlie et al. (2001), Huang et al. (2003), Hedenfalk et al. (2001), and Farmer et al. (2005)) were normalized according to platform type. In particular, Affymetrix platforms (the `Farmer` and `Huang` datasets) were normalized by rma, whereas cDNA platforms (`Sorlie` and `Hedenfalk`) were normalized using the methods described in Smyth and Speed (2003) and implemented in the

R package *limma*. Following normalization, platform-specific annotations were mapped to Entrez gene identifiers and the resulting lists merged to obtain a set of 2064 genes. The datasets are availabile as a R object in the R package *xdeBreastData* (available from author).

6 Validation

This section is comprised of two parts. In the first, we simulate differential expression in a subset of the genes for three lung cancer studies. As the set of genes that are differentially expressed are known through simulation, we assess performance of our model relative to alternative approaches using diagnostics such as the area under the ROC curve (AUC). In the second part of this section, we evaluate the shrinkage properties of the Bayesian model by applying *XDE* to multiple splits of a single study. Comparisons of *XDE* to alternative methods for cross-platform analysis are discussed throughout.

6.1 Simulated data

Our simulations are based on three publicly available lung cancer datasets that we refer to by institution: `Harvard` (Bhattacharjee et al., 2001) , `Michigan` (Beer et al., 2002), and `Stanford` (Garber et al., 2001). See Section 5 for a brief description of these datasets.

Simulation algorithm. We begin by describing an approach for generating artificial datasets for which the true set of differentially expressed genes is known. We append the superscript '$\star$' to parameters used in the simulation to distinguish the *true* values from the corresponding variables in the Bayesian model.

The simulation uses only stage I or II adenocarcinomas in the `Harvard` (n=83), `Stanford` (n=11), and `Michigan` (n=61) studies. Late stage adenocarcinomas were excluded as the heterogeneity of these tumors is likely to be much greater. Letting S denote the total number of samples for each study in the simulated dataset, we randomly assigned the clinical variable $\psi^\star = 0$ to half of the samples and $\psi^\star = 1$ to the remaining half. Although there is non-trivial heterogeneity within the adenocarcinomas, these differences become small (on average) after random assignment into classes and provide a background noise that would be difficult to simulate *de novo*. Independently for each

gene, we simulate $\delta_g^\star = 1$ from a Bernoulli with some probability $\xi^\star$ that is common to all genes, and set $\delta_g^\star = 0$ otherwise. For genes with $\delta_g^\star = 1$ we thereafter generate "true" offsets $(\Delta_{g1}^\star, \Delta_{g2}^\star, \Delta_{g3}^\star)$ from a multivariate normal distribution

$$\begin{bmatrix} \Delta_{g1}^\star \\ \Delta_{g2}^\star \\ \Delta_{g3}^\star \end{bmatrix} \sim \mathrm{N}\left(k^\star \begin{bmatrix} s_{g1} \\ s_{g2} \\ s_{g3} \end{bmatrix}, \frac{1}{c^\star} \begin{bmatrix} s_{g1}^2 & r_1^\star s_{g1} s_{g2} & r_2^\star s_{g1} s_{g3} \\ r_1^\star s_{g2} s_{g1} & s_{g2}^2 & r_3^\star s_{g2} s_{g3} \\ r_2^\star s_{g3} s_{g1} & r_3^\star s_{g3} s_{g2} + & s_{g3}^2 \end{bmatrix} \right), \tag{11}$$

where s_{g1}, s_{g2} and s_{g3} are the empirical standard deviations for the adenocarcinoma samples in `Harvard`, `Michigan`, and `Stanford`, respectively, and $k^\star$, $c^\star$ and $\mathbf{r}^\star$ are parameters in the simulation procedure. Letting x_{gsp} denote the original adenocarcinoma expression values, we generate the corresponding artificial data as

$$x_{gsp}^\star = \begin{cases} x_{gsp} + (2\psi_{sp}^\star - 1)\Delta_{sp}^\star & \text{if } \delta_g^\star = 1, \\ x_{gsp} & \text{otherwise.} \end{cases} \tag{12}$$

We consider a gene g as differentially expressed if $\delta_g^\star = 1$. Differential expression is concordant if $\mathbf{\Delta}_g^\star$ have the same sign in all studies and discordant if $\mathbf{\Delta}_g^\star$ have opposing signs. Concordant and discordant differential expression are special cases of differential expression that we consider separately. Note that the simulation parameters $\mathbf{r}^\star$, $c^\star$, and $k^\star$ control the proportion of differentially expressed genes that are concordant in the simulation. For instance, increasing $\mathbf{r}^\star$ and $c^\star$ have the effect of increasing the percentage of concordantly differentially expressed genes. See Table 1 for a complete listing of the simulation settings evaluated.

For each of the Simulations A-R in Table 1, we develop summary measures, referred to as *scores*, to quantify concordant ($\mathcal{C}$), discordant ($\mathcal{D}$), or the union of (differentially) expressed ($\mathcal{E}$) genes. Section 4.2 discusses the summary statistics proposed for the Bayesian model, as well as alternative methods for summarizing differential expression. We emphasize that for a gene g, $\mathcal{C}_g$, $\mathcal{D}_g$, and $\mathcal{E}_g$ are defined on a set of studies, as opposed to differential expression in a single study. To illustrate, recall that in our simulations, a gene g is differentially expressed in all studies when the indicator $\delta_g^\star = 1$. Table 2 illustrates the possible patterns of differential expression for $P = 2$ studies and $G = 4$ genes.

Simulation	$k^\star$	S	$c^\star$	$\mathbf{r}^\star$	$\xi^\star$
A†	0.5	4	0.5	(0.1, 0.2, 0.4)	0.10
B	·	·	·	·	0.50
C	·	·	·	(0.8, 0.9, 0.92)	0.10
D	·	·	·	·	0.50
E†	·	8	0.5	(0.1, 0.2, 0.4)	0.10
F	·	·	1	·	0.10
G	·	·	·	·	0.50
H	·	·	·	(0.8, 0.9, 0.92)	0.10
I	·	·	·	·	0.50

Simulation	$k^\star$	S	$c^\star$	$\mathbf{r}^\star$	$\xi^\star$
J†	0	16	10	(0.1, 0.2, 0.4)	0.10
K	·	·	·	·	0.50
L	·	·	·	(0.8, 0.9, 0.92)	0.10
M	·	·	·	·	0.50
O	·	32	20	(0.1, 0.2, 0.4)	0.10
P	·	·	·	·	0.50
Q	·	·	·	(0.8, 0.9, 0.92)	0.10
R	·	·	·	·	0.50

Table 1: Each row in the table displays parameters used to simulate an artificial dataset of three studies with S samples in each. From Equation 11, $k^\star$ and $c^\star$ control the location and scale of the simulated Gaussian offsets, respectively. Together, $\mathbf{r}^\star$, $k^\star$, and $c^\star$ control the degree of concordance of the simulated offsets. The probability that a gene was differentially expressed was $\xi^\star$. † Ten artificial datasets were generated from one set of simulation parameters $(S, k^\star, c^\star, r^\star, \xi^\star)$ but with different seeds for the random number generator. By varying only the seed for the random number generator, we can assess the sensitivity of performance measures such as AUC to randomly generated quantities in the simulation (e.g., the set of genes with $\delta^\star = 1$). In general, simulation parameters were selected such that the AUC statistic from the simulations ranged between 0.6 and 0.9.

gene	$\delta^\star$	sign $(\mathbf{\Delta}^\star)$	$\mathcal{E}$	$\mathcal{C}$	$\mathcal{D}$
1	0	·	0	0	0
2	1	$\{-,-\}$	1	1	0
3	1	$\{-,+\}$	1	0	1
4	1	$\{+,+\}$	1	1	0

Table 2: A trivial example of a dataset with four genes and two studies. For each gene, we evaluate three possible truths for differential expression defined over the set of studies: concordant differential expression in both studies ($\mathcal{C}_g = 1$), discordant differential expression ($\mathcal{D}_g = 1$), and differential expression that is either concordant or discordant across studies ($\mathcal{E}_g = 1$).

Evaluation procedures

Let $u_*(g)$ denote any of the scores defined above for a gene g. If, for a fixed threshold $a > 0$, we classify each g as being (overall, concordantly or discordantly) differentially expressed if $u_*(g) > a$, then we obtain a standard two-by-two table containing the number of false negatives FN(a), false positives FP(a), true negatives TN(a), and true positives TP(a). For example, the number of true negatives is given by

$$\text{TN}(a) = \sum_{g=1}^{G} I(u_*(g) \leq a \text{ and } \delta_g^\star = 0), \tag{13}$$

and the remaining entries of the table are defined in a similar fashion. The false positive and true positive rates associated with the statistics U_g and threshold a are then

$$\mathrm{FPR}(a) = \frac{\mathrm{FP}(a)}{\mathrm{TN}(a) + \mathrm{FP}(a)} \quad \text{and} \quad \mathrm{TPR}(a) = \frac{\mathrm{TP}(a)}{\mathrm{FN}(a) + \mathrm{TP}(a)}, \tag{14}$$

respectively. Plotting $\mathrm{FPR}(a)$ against $\mathrm{TPR}(a)$ as a varies produces the standard receiver operating characteristic (ROC) curve associated with the statistics U_g. The area under the ROC curve, AUC, is a nonparametric measure of the quality of the statistic, with values close to unity (*i.e.*, a statistic that simultaneously achieves FPR close to zero and TPR close to one) being the best.

As an alternative to ROC curves, which are based on false and true positive rates, we also considered the false discovery rate (FDR) of the statistics $u_*(g)$ as a function of the number of genes determined to be differentially expressed. Specifically, for each threshold a, we plotted the number of discoveries, $\sum_{g=1}^{G} I(u_*(g) > a)$, against

$$\mathrm{FDR}(a) = \frac{\mathrm{FP}(a)}{\mathrm{FP}(a) + \mathrm{TP}(a)}.$$

As expected, the FDR increases as the number of overall discoveries increases. Curves close to the horizontal axis are preferable to those having a more rapid increase of FDR with the number of discoveries. Similarly, we plotted the number of non-differentially expressed genes, $\sum_{g=1}^{G} I(u_*(g) < a)$, against the missed discovery rate

$$\mathrm{MDR} = \frac{\mathrm{TN}(a)}{\mathrm{FN}(a) + \mathrm{TN}(a)}.$$

Again, curves close to the horizontal axis are preferable to those having a more rapid increase of MDR with the number of negative discoveries.

Simulation results

We generated artificial datasets for Simulations A - R in Table 1 as described previously. Initial model parameter values for *XDE* were chosen to specify little prior knowledge: $\alpha_a = \beta_a = \alpha_b = \beta_b = 1$, $p_a^0 = p_a^1 = p_b^0 = p_b^1 = 0.1$, $\nu_\rho = \nu_r = 4$ and $\alpha_\xi = \beta_\xi = 1$. The values for the tuning parameters in the Metropolis–Hastings algorithm were chosen to achieve a robust algorithm, not

to optimize convergence and mixing properties for this particular data set. In all updates of type (iii) we used $\varepsilon = 0.01$. For updates of type (iv) we used $\varepsilon = 0.5$ in updating σ^2_{gp} and φ_{gp}, and $\varepsilon = 0.1$ in updating l_p and λ_p. In updates of types (v), (vi) and (vii) we used $\varepsilon = 0.1$, $\varepsilon = 0.05$ and $\varepsilon = 0.02$, respectively.

To monitor convergence and mixing properties, we inspected trace plots of the various simulated variables, as in Supplementary Figure 9. We observed that most parameters converge relatively quickly and that the model parameters coincide in many cases with the true values in the simulation. For instance, 10% of the genes were simulated to be differentially expressed in Simulation A ($\xi^* = 0.10$) and traceplots of the ξ parameter in the Bayesian model show that this parameter has converged to a value near 0.12.

For each simulation, performance of the cross-study scores were assessed by the AUC, FDR, and MDR criteria. A graphical display of the results for Simulation A is shown in Figure 2. The Bayesian model has a higher AUC (panel 1), as well as a lower FDR and MDR than the alternative scores over a range of cut-offs for evaluating $\mathcal{C}$ (panels 2 and 3, respectively). Figure 3 plots the AUC for $\mathcal{C}$ in Simulations A-R. The corresponding AUC statistics for $\mathcal{E}$ and $\mathcal{D}$ are provided in Supplementary Figures 10(a) and 10(b). The `t`-score generally does worse than the other methods for small samples sizes. This is likely a result of inflated signal to noise ratios in genes with very small variation across the four samples. That `SAM` is notably better than the `t` in such situations is consistent with this observation.

The different seeds used in the simulations produce different artificial datasets as both the δ^* and Δ^* are randomly generated. To assess the sensitivity of the AUC to the seed used for random number generation, Figure 4 plots the AUC for datasets simulated from the same simulation parameters, differing only by seed.

Because of the extensive nature of the simulations, we visually assess the relative performance of the Bayesian method to alternative approaches in scatterplots of the AUC (Figures 3 and 4). Points beneath the identity line are simulations in which the Bayesian score had a higher AUC than an alternative method evaluated on the same dataset. In general, the Bayesian model outperforms the four alternative methods for cross-study analysis of differential gene expression across a range of simulated parameters (Figure 3). Our overall assessment does not appear to be sensitive to the random quantities simulated in these datasets (Figure 4). As the sample sizes of the indi-

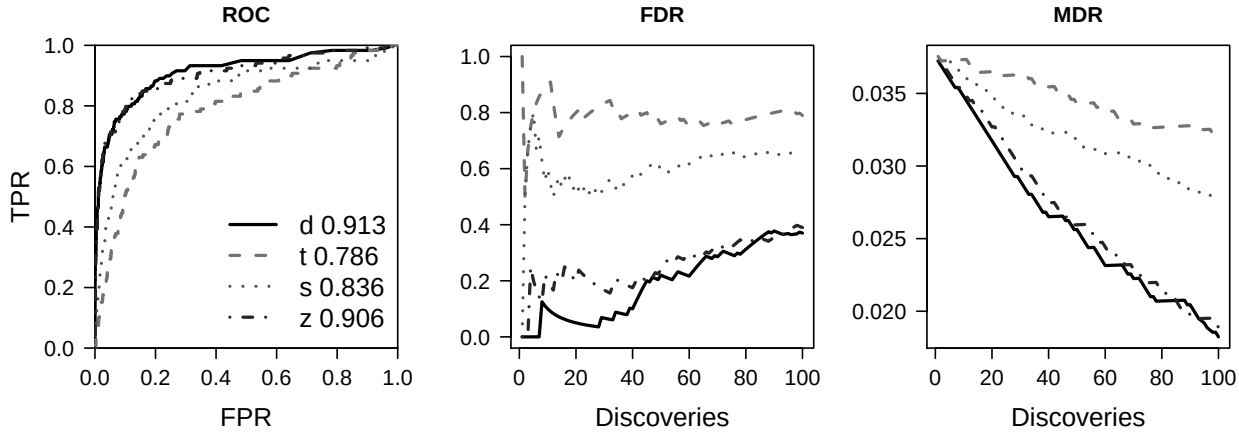


Figure 2: Performance diagnostics for scores quantifying $\mathcal{C}$ in Simulation A. The letter d in the legend corresponds to the Bayesian score. Although the `SAM`-score does markedly better than the `t`-score when the the individual studies are small (S = 4), considerable improvement can be obtained by a more formal borrowing of strength across studies in the Bayesian and `z`-scores.

vidual studies increase, the relative benefit of borrowing strength across genes and studies in the hierarchical model diminishes. Instances in which the `z`-score has a better AUC than the corresponding Bayesian statistic (e.g., panel[2, 1] in Figure 3), most often occurred when the simulated data was particularly noisy and the AUC from all methods were at the low-end of the range. In such instances, scatterplots of the study-specific effect sizes were largely uncorrelated (data not shown). Scatterplots of a study-specific statistic for effect-size, such as `t`, may be a useful indicator of whether the Bayesian model is likely to improve on simpler alternatives. In instances where the data is negatively correlated across studies, this may induce the 'wrong' borrowing of strength. In simulated datasets with no signal (data not shown), gene-specific posterior probabilities in the Bayesian model were approximately zero.

6.2 Split study validation

To assess the baseline behavior of *XDE*, we split the `Huang` study into four disjoint parts, treating each part as an independent study. We randomly assigned 5 estrogen receptor (ER) negative and 16 ER positive samples to each split. Split study validation has been used by others to assess meta-analytic methodologies for gene expression analysis. In particular, Gentleman et al. (2005) use split study validation to illustrate their implementation of the cross-platform statistic introduced by Choi et al. (2003). In this simplified setting, we avoid the potential difficulties of

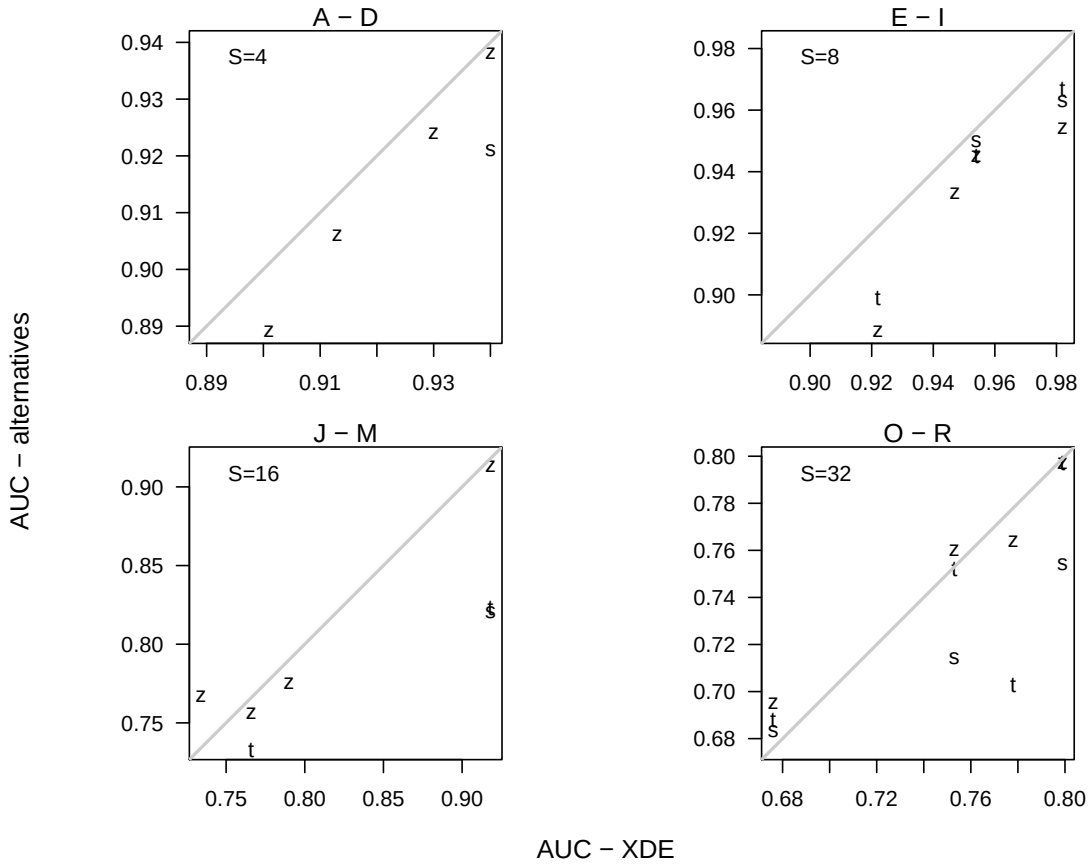


Figure 3: The AUC for concordant differential expression in Simulations A - D (top left), E - I (top right), J - M (bottom left), and O - R (bottom right) was calculated for each of the alternative methods (`t`, `SAM`, and `z`) and plotted against the AUC obtained from the Bayesian model. The diagonal line in each panel is the identity. The lower limit for the axes are based on the minimum of the AUC's from the Bayesian and `z`-scores; hence the `t` and `SAM` scores are not always plotted. See Table 1 for the simulation parameters.

cross-platform analyses that can arise from technological and/or biological differences between studies. For instance, differences in the annotation of the probes or ethnic composition of the study populations may each contribute to discrepant results in a meta-analysis, but such concerns are reduced when splitting a single study.

After fitting the Bayesian model to the four splits, traceplots for the parameters a, b, l, t, γ^2, c^2, τ^2, ξ, ρ, and r (each of which are updated by Metropolis-Hastings proposals) were used to evaluate convergence (see Supplementary Figure 11). We define the Bayesian effect size `BES` for gene g and platform p, by $\frac{\delta_g \Delta_{gp}}{c\tau_p \sigma_{gp}^{b_p}}$, and use this as a study-specific Bayesian estimate of differential expression, contrasting it with the `z`, `t`, and `SAM` statistics. Scatterplots of the study-specific `t`-, `z`-,

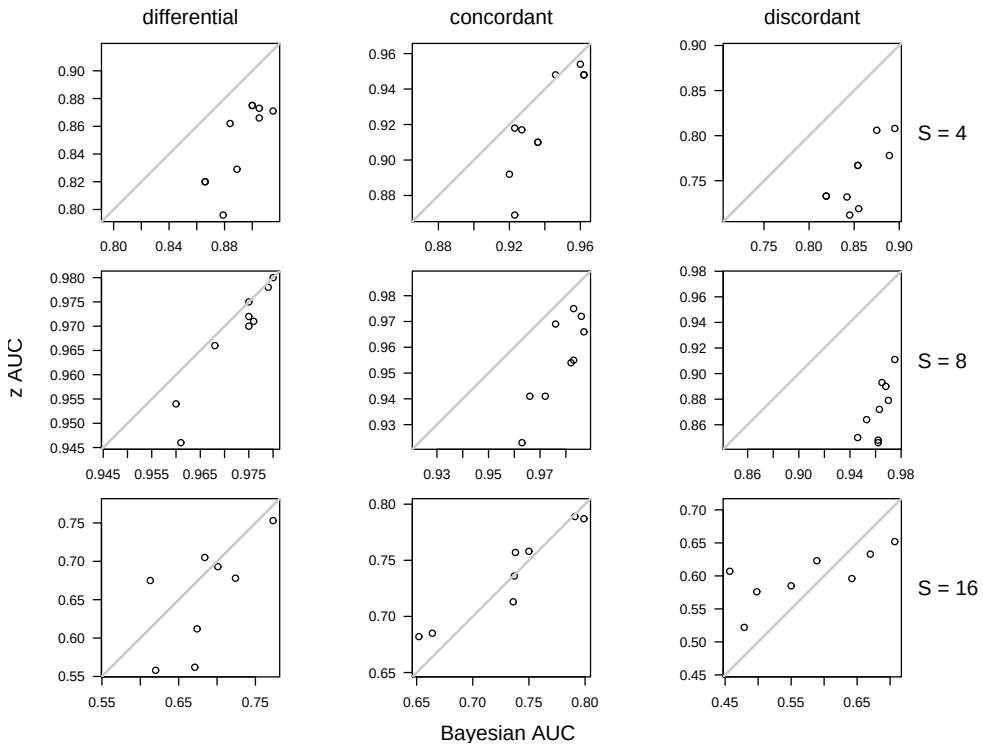


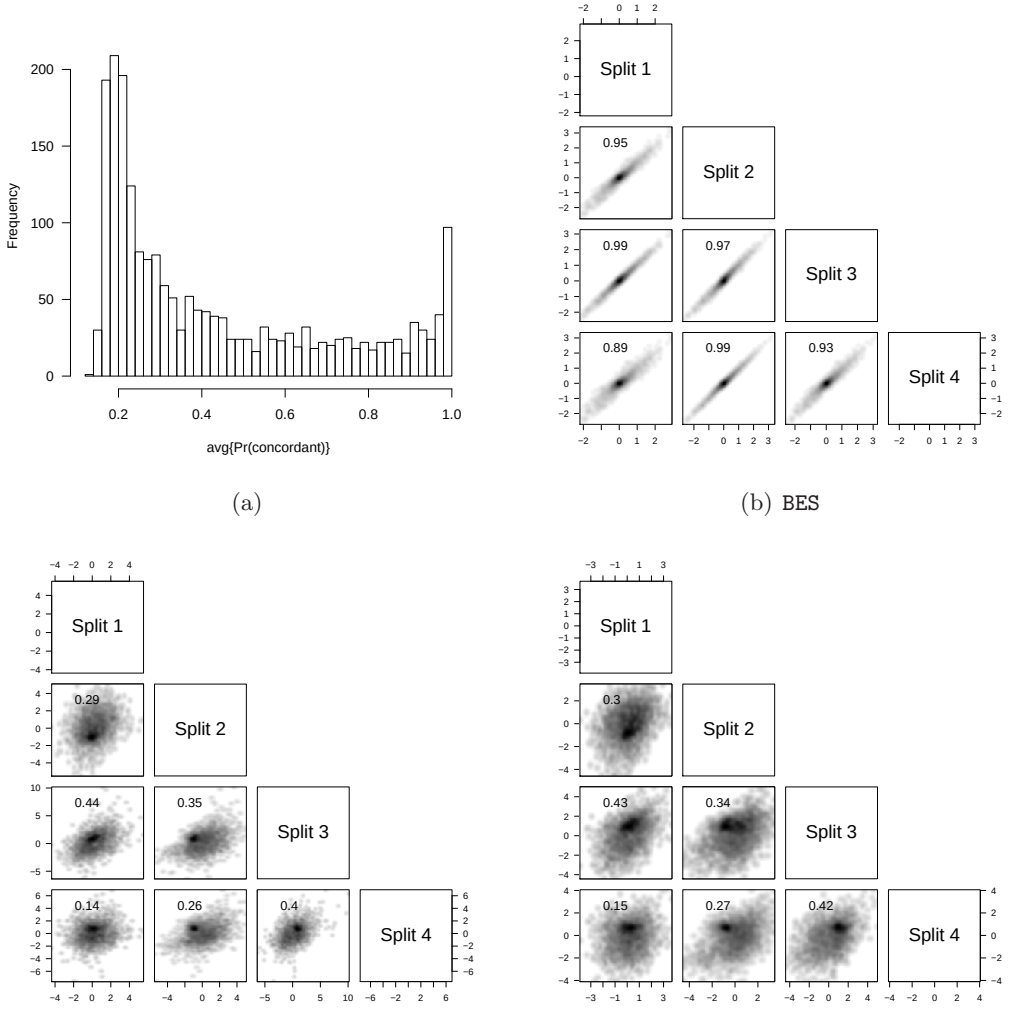
Figure 4: To evaluate the extent to which random draws of δ^* and $\mathbf{\Delta}^*$ influence the performance of the different methods, we simulated 10 artificial datasets for Simulation $A^\dagger$ (top row), $E^\dagger$ (row 2), and $J^\dagger$ (row 3) using different seeds for the random number generator. In each panel, we plot the AUC from the Bayesian model (horizontal axis) against the AUC from the `z`-score (vertical axis). Not shown are the AUC from the `SAM`- and `t`-scores. The columns depict the three different ways to evaluate differential expression. Posterior averages for the Bayesian statistic (Section 4) were calculated from 1000 iterations (saving every 20^{th} iteration of 20,000 iterations) following a burnin of 2000 iterations.

and `BES` statistics are shown in Figure 5. If we consider the `t`, `SAM` (not shown), and `z` statistics as evidence of differential expression in a single study, we observe that the evidence is study-dependent with only moderate correlation of these statistics across the splits (Figures 5(c) and 5(d)). Hence, scatterplots of the study-specific statistics provide two important pieces of information: first, even in a scenario that minimizes inter-study discordance, the variation across studies of the effect size statistics underscore the difficulty of identifying genes that show consistent evidence of differential expression; secondly, while the scatterplots do not lend themselves directly to identifying a list of genes for follow-up, the moderate correlation among the study-specific statistics does motivate an approach that uses the information from all of the studies.

A set of concordant differentially expressed genes emerges from the visualization of the `BES` scatterplots in Figure 5(b). Through modeling the inter-relationships of genes and studies at higher

levels of the model, the Bayesian model shrinks noisy genes to zero without requiring extensive filtering prior to the analysis. The cigar-shaped pattern in Figure 5(b) is typical when fitting the Bayesian model, though the correlation is higher than what one may expect to observe when the studies are independent and use different platforms (see Section 7). In choosing a list of genes to follow for subsequent laboratory investigation, the $\mathrm{PM}_{\mathrm{e}}(g)$, displayed in Figure 5(a), can be used to rank the evidence of concordant differential expression.

Validation of microarray experiments typically involves assaying RNA transcript abundance of candidate genes selected from the high-throughput technologies by low-throughput platforms, such as qRT-PCR. As the $\mathrm{PM}_{\mathrm{e}}(g)$ identifies genes whose differential expression is relatively study- and/or platform-independent, validation of the gene list selected by $\mathrm{PM}_{\mathrm{e}}(g)$ may be less likely to result in false discoveries as suggested by the simulations in the previous section. Hence, in addition to increasing the power to detect differentially expressed genes in the context of small sample size, meta-analysis could potentially result in a cost-savings downstream of the analysis. However, the more likely scenario with meta-analysis is the attempt by an impartial investigator not directly associated with the primary studies to synthesize the information. Reporting the genes and pathways that are affected in each of the studies, as well as the genes and pathways that appear discordantly regulated are important. Whether the goal is to produce a gene list that is likely to be validated by other platforms, or to explore more deeply the biological explanations of concordance and discordance, the Bayesian model provides a useful mechanism for achieving these goals.



(a) (b) BES

(c) t-statistics (d) z-scores

Figure 5: Top left: Distribution of the posterior probability for concordant differential expression, $\mathrm{PM}_{\mathrm{e}}(g)$. Panels 2-4 are scatterplots of study-specific measures of differential expression in the split-study validation. `t` and `z` statistics, estimated independently for each study, show considerable variation across studies with discordance that is probably within the noise of the experiment. The modest correlation of the study-specific statistics motivates an approach that more effectively models the inter-study and inter-gene relationships. The `BES` (top right) shows how noisy genes are shrunk towards zero, whereas genes in quadrants (+, +) and (-, -) that show some evidence of differential expression in each of the studies are shrunk less. As the $\mathrm{PM}_{\mathrm{e}}(g)$ is useful for ranking concordant differential expression in multiple studies or platforms, the highest ranked genes are typically genes whose differential expression was not platform- or study-dependent. As the goal of many microarray experiments is to select genes for subsequent validation by other platforms for measuring transcript abundance (such as qRT-PCR), a ranking that is not platform- and study-dependent may facilitate this effort.

7 Experimental data example

Estrogen receptor is an important risk factor for breast cancer tumorigenesis and several gene expression studies have collected phenotypic information on estrogen receptor (ER) status (positive or negative). In this section, we fit the Bayesian model to four publicly available datasets described in Huang et al. (2003) (`Huang`), Hedenfalk et al. (2001) (`Hedenfalk`), Farmer et al. (2005) (`Farmer`), and Sorlie et al. (2001) (`Sorlie`), using ER status as the clinical variable. Table 3 shows the distribution of ER for the four breast cancer studies. The main purpose of the integration of the breast cancer studies is to define a profile of differentially expressed genes using our Bayesian hierarchical model that is less likely to be platform-dependent. See Section 5 for a brief description of the data and pre-processing steps. Because the studies involve different gene expression platforms, we cross-reference the study-specific gene annotations by Entrez gene identifiers and focus our discussion on the set of 2064 Entrez genes that were present in each of the four studies.

	platform	ER-	ER+
Hedenfalk	cDNA	6	10
Sorlie	cDNA	30	81
Farmer	Affymetrix hu133a	22	27
Huang	Affymetrix hu95av2	23	65

Table 3: Distribution of the estrogen receptor in the three studies

When fitting the Bayesian hierarchical model to the breast cancer datasets, we found it unnecessary to change the hyperparameters and tuning parameters for the Metropolis–Hastings algorithm from their default values (see Table 4). To monitor the convergence and mixing properties of the Markov chain, we used visual inspection of trace plots of the various variables simulated. The slowest convergence and mixing properties occurs for the four hyper-parameters θ_p, λ_p, t_p and l_p, see for example the trace plots of $l_p, p = 1, 2, 3$ in Supplemental Figure 12. A burn-in of 5000 iterations is sufficient for convergence in most instances, but this should be evaluated on a case by case basis. We calculated posterior statistics using every 20th iteration after 2000 iterations burnin.

The histograms in Figure 6 display $\mathrm{PM}_{\mathcal{E}}(g)$, (top), $\mathrm{PM}_{\mathcal{C}}(g)$ (middle), and $\mathrm{PM}_{\mathcal{D}}(g)$ (bottom). Among the 77% of genes that are differentially expressed, 42% are concordant and 35% are discordant. The Bayesian model likely overestimates the true proportion of genes that are differentially expressed. This is a consequence of the model putting a very small variance on the Δ parameter.

In particular, c^2 becomes small. In our formulation, differential expression (whether concordant or discordant) is defined as a departure of any magnitude from $\Delta = 0$. The model is likely sensitive to systematic artifacts in the data that are confounded with phenotype. Nevertheless, the overall ranking of the genes and the main conclusions of the analysis are not affected. Estimates of concordant differential expression, the most common goal of most integration efforts, appear unaffected by the ξ inflation– the posterior expected proportion of false positives for the experimental data were low for a range of reasonable cutoffs for differential expression. In particular, the posterior expected proportion of false positives using thresholds of 0.5 and 0.9 for $\mathrm{PM}_{\mathcal{C}}(g)$ ranges from 0.22 to 0.04, respectively (Efron and Tibshirani, 2002). The software implementation of the Bayesian model flags instances of small c^2 alerting the user that the posterior averages may be miscalibrated. Evaluating different priors for Δ is a future direction of this research.

We explore concordant and discordant differential expression separately, combining visualizations that are effective for summarizing the overall reproducibility (pairwise scatterplots of effect size) with statistics from the Bayesian model that can be used to target a specific subgroup of genes that appear to be concordantly (Figure 7) or discordantly (Figure 8) regulated in the different studies. Genes in the highest decile of $\mathrm{PM}_{\mathcal{C}}(g)$ (to the right of the vertical dashed line in Figure 7(a)) are plotted with a different symbol (circles) and color (black) in the pairwise scatterplots of `BES`, `t`-, and `z`-statistics. The Bayesian model shrinks noisy estimates of the effect size towards zero (panel b, Figure 7(b)), whereas consistent estimates of differential expression in the studies are shrunk less and appear in quadrants (-, -) and (+, +) of the pairwise scatterplots of `BES` in panel b.

Figure 8 explores discordance. Panels b - d are the same as in Figure 7, but with an emphasis on inter-study discordance identified by thresholding the upper 5% of genes by the $\mathrm{PM}_{\mathcal{D}}(g)$ (genes to the right of the vertical dashed line in Figure 8(b)). Again, emphasis is placed on a subset of genes through different plotting symbols ("x") and color (black). Note that almost all of the discordance in the scatterplots shown in Figure 8b - d arise from pairwise comparisons of cDNA platforms (`Sorlie` and `Hedenfalk`) to the Affymetrix platforms (`Farmer` and `Huang`). Discordance between Affymetrix and cDNA platforms may arise, for instance, as a result of probes hybridizing to different transcripts of the same gene. Note that the scatterplots comparing like platforms (`Sorlie` versus `Hedenfalk` (both cDNA) and `Farmer` versus `Huang` (both Affymetrix)), the effect size

estimates of the highlighted genes are positively correlated. Exploring the discrepancies across the different platforms is an interesting future direction of this research.

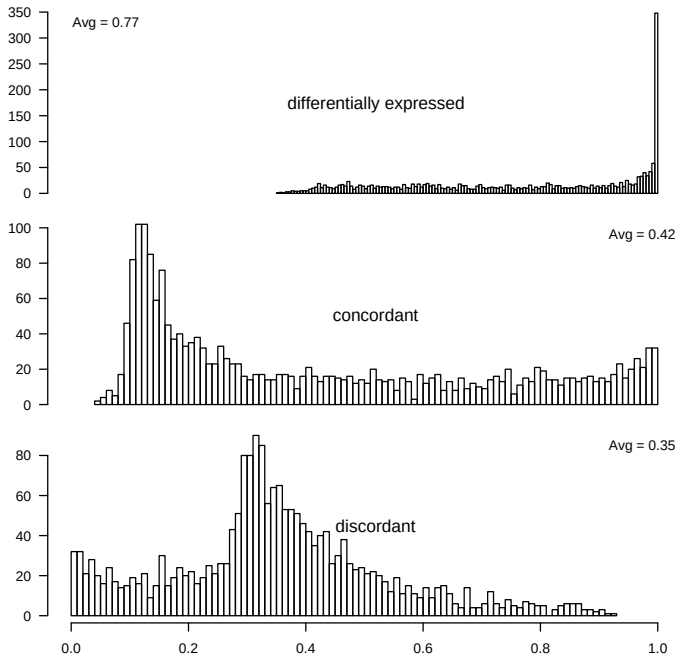
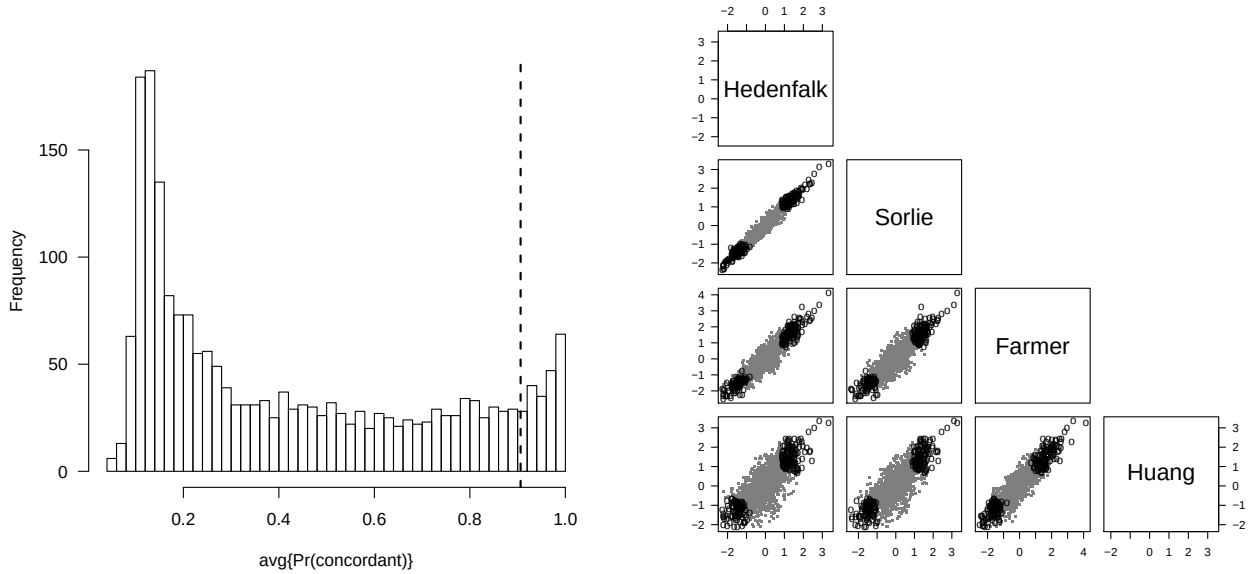
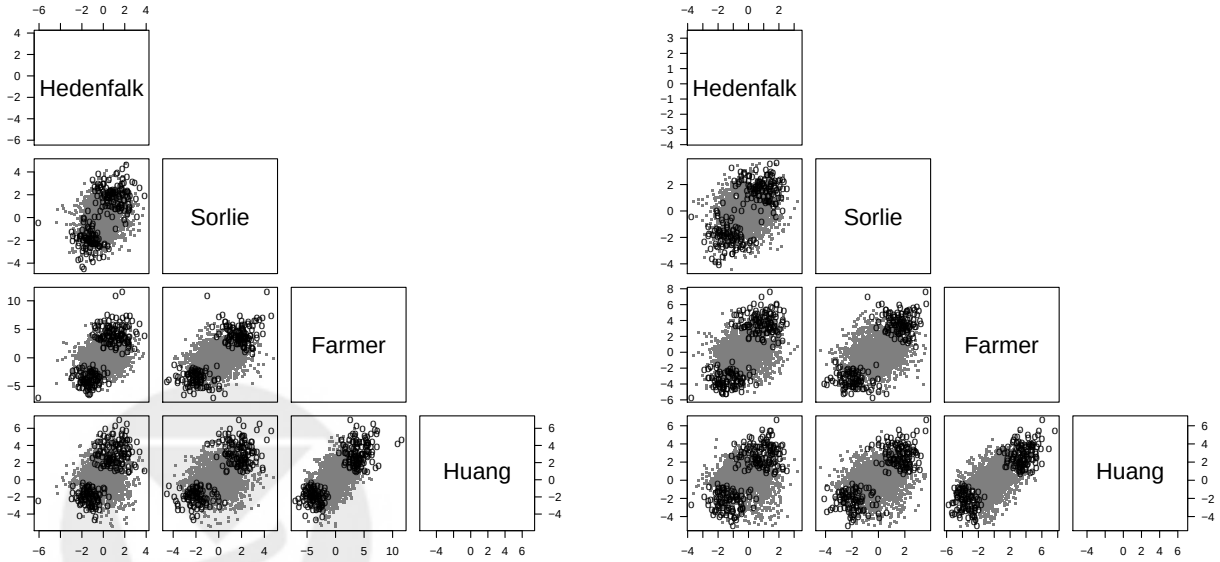


Figure 6: Histograms of gene-specific posterior probabilities for differential expression $\{\mathrm{PM}_{\mathcal{E}}(g)\}$, concordant differential expression $\{\mathrm{PM}_{\mathcal{C}}(g)\}$, and discordant differential expression $\{\mathrm{PM}_{\mathcal{D}}(g)\}$. The top figure suggests that a high proportion of the genes (77%) show some evidence of differential expression in one of more studies. In the analysis of multiple studies, the differential expression can be further classified as concordant (35%) or discordant (42%). We emphasize the utility of the ranks of these statistics for identifying genes whose differential expression is consistent (rank of $\mathrm{PM}_{\mathcal{C}}(g)$) or inconsistent (rank of $\mathrm{PM}_{\mathcal{D}}(g)$) across studies.



(a) The vertical dashed line is drawn at the 90th percentile.

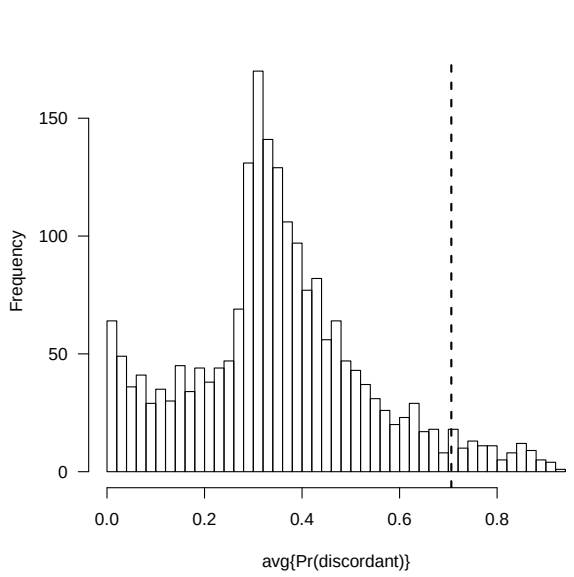
(b) `BES`



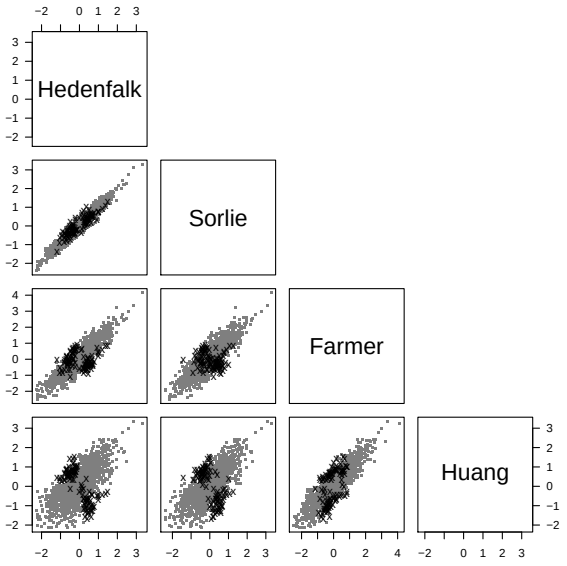
(c) `t`

(d) `z`-score

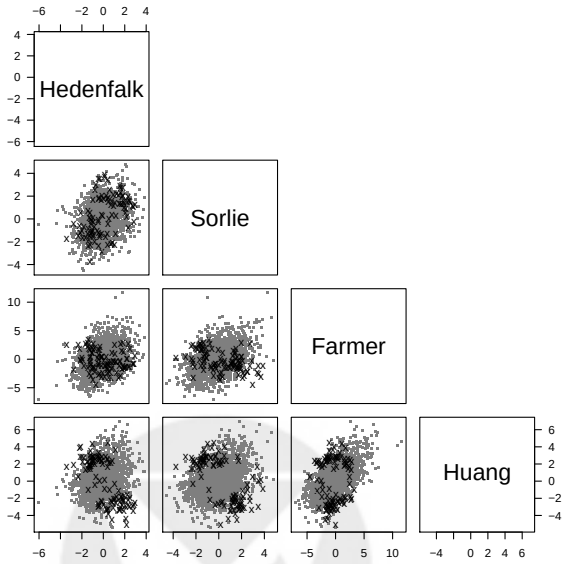
Figure 7: Ranking genes by the $\mathrm{PM}_{\mathrm{e}}(g)$ is useful for exploring inter-study agreement of differential expression. Here, we threshold genes by the 90^{th} percentile of the distribution for the posterior average of concordant differential expression, $\mathrm{PM}_{\mathrm{e}}(g)$ (panel a). Pairwise scatterplots of the study-specific statistics for the four breast cancer studies are provided in panels b - d. A different plotting symbol (circles) and color (black) is used for the genes in the highest decile of $\mathrm{PM}_{\mathrm{e}}(g)$. The posterior expected proportion of false positives corresponding to this threshold is approximately 0.04.



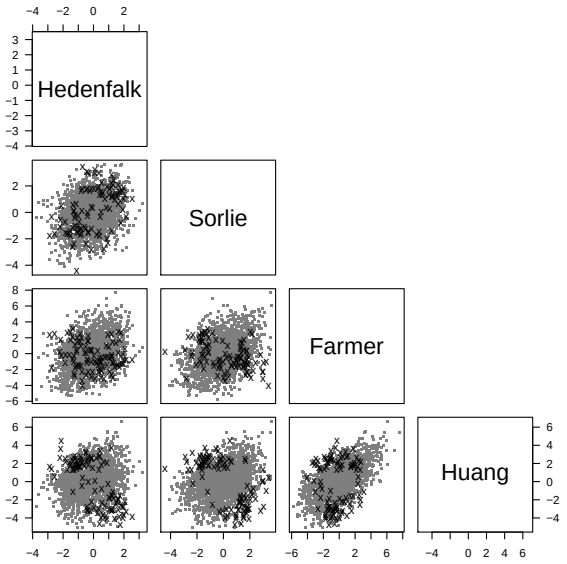
(a) The vertical dashed line is drawn at the 90th percentile.



(b) `BES`



(c) `t`



(d) `z`-score

Figure 8: The posterior average of the probability of discordant differential expression, $\mathrm{PM}_{\mathcal{D}}(g)$ (panel a), can be used to explore discordance. Here we threshold pairwise scatterplots of the study-specific statistics from the four breast cancer studies (panels c and d) by the 95^{th} percentile of the $\mathrm{PM}_{\mathcal{D}}(g)$ distribution (panel a). Again, we use a different plotting symbol (x) and color (black) for genes surpassing this threshold to emphasize the discordance. In particular, note that almost all of the discordance in the scatterplots of panels b - d arise from pairwise comparisons of cDNA platforms (`Sorlie` and `Hedenfalk`) to the Affymetrix platforms (`Farmer` and `Huang`). For the two scatterplots comparing more similar platforms (`Sorlie` versus `Hedenfalk` and `Farmer` versus `Huang`), the effect size estimates of the highlighted genes are positively correlated.

8 Closing remarks

In this paper we define a hierarchical Bayesian model for microarray expression data collected from several studies and use it to identify genes that show differential expression between two conditions. We evaluated its performance using artificial data, real data, and a novel "split-sample" validation approach that provides a model-free assessment of the behavior of the model not only under the null hypothesis but also under a realistic alternative. The simulation results from the artificial data demonstrate the advantages of a Bayesian model. Compared to a more direct combination of t- or SAM-statistics, the $1 - \text{AUC}$ values for the Bayesian model are roughly half of the corresponding values for the t- and SAM-statistics. Furthermore, the simulations provide guidelines for when the Bayesian model is most likely to be useful. In small studies the Bayesian model generally outperforms other methods when evaluated by AUC, FDR, and MDR across a range of simulation parameters, and these differences diminish for larger sample sizes in the individual studies. The split-study validation illustrates appropriate shrinkage of the Bayesian model in the absence of platform-, sample-, and annotation-differences that otherwise complicate experimental data analyses. Using experimental data from four high-thoughput gene expression studies for breast cancer, we evaluate differential expression using estrogen receptor status (positive or negative) as the clinical covariate of interest. We evaluate concordant and discordant differential expression separately, using posterior averages from the Bayesian model to identify subsets of genes that may be partly explored in-silico. For instance, Figure 8 identified a subset of genes in the breast studies that were discordant across platforms (cDNA versus Affymetrix) but remain positively correlated within a platform (cDNA versus cDNA and Affymetrix versus Affymetrix). Further filtering this list by genes with known alternative transcripts that may be measured differently by platform (probes hybridizing to different splicoforms) may add insight to the differential expression of splicoforms of a gene. Because such in-silico hypotheses can only be validated by laboratory based methods such as qRT-PCR, we leave this as an open thread for future investigation.

Of the models previously proposed in the literature, the model of Conlon et al. (2006) is conceptually closest to ours. The Conlon model is designed for cross-study within-platform analyses and was not directly applicable to the case studies in our article. However, it is useful to contrast the technical features of the two approaches. Both are hierarchical Bayesian models, and both have

a differential expression indicator for each gene. Differences emerge in how each model handles the increased variation in expression values for differentially expressed genes. We assign separate distributions to the expression values of samples in each condition. Conlon et al. (2006) assume that the expression values for differentially expressed genes are independent, but with an increased variance; they do not make use of the condition information for each sample. In addition, we adopt a more refined and flexible model for the covariance structure of the expression values. A practical consequence of these differences is in the application of the models to gene expression data from different platforms. In particular, our model can be fit to multiple studies regardless of platform, whereas Conlon et al. (2006) is most applicable for combining replicates from a single platform. In this sense, the Conlon et al. (2006) model could be viewed as a special case of our model when the technological differences in scale and variation are approximately zero and conjugacy between location and scale is true.

Our hierarchical model does not require that studies be measured on the same platform and this generality has advantages and disadvantages. One advantage is that we model the differences in scale and variation of expression intensities across platforms directly, removing some of the need for extensive normalization and nonparametric rank-based approaches. However, in any multi-study analysis, discordance can arise from biological differences in the sample populations of each study, as well as technological effects related to the design and implementation of specific array technologies. Modeling gene expression in a hierarchical way, we borrow strength across studies and genes, by shrinking noisy estimates to zero and capturing correlated signals from the different studies. If the discordant signal is stronger than the concordant signal, this may induce a 'wrong' borrowing of strength in which concordant differential expression is seen as noise and shrunk towards zero. Simple scatterplots of study-specific measure of effect size, such as the `SAM` statistic, are a very simple diagnostic that can be performed before fitting the Bayesian model. Typically, one may see a cloud of effect size statistics near zero and some correlation in the positive (+, +) and negative (-, -) quadrants. In such situations, the Bayesian model will shrink the cloud to zero, providing less shrinkage of the concordantly differentially expressed genes and more shrinkage of the discordant differentially expressed genes.

It is common in the analysis of high-throughput gene expression data to apply a gene-selection procedure prior to the formal analysis of differential expression. For instance, when estimating

differential expression by a statistic that has in its denominator an estimate of the across-sample variation, one may wish to remove genes of low abundance that show very low across-sample variation. In our Bayesian model, each gene has a parameter representing the numerical value of its differential expression. The priors for these parameters have a point mass at zero, corresponding to no differential expression. Such a parameterization removes the need to apply gene selection techniques prior to the analysis of differential expression, permitting a more direct and model-based procedure. See also the discussion in Ishwaran and Rao (2003, 2005).

When fitting the Bayesian model to pure noise, the model behaves appropriately and the estimated proportion of differentially expressed genes (the union of concordant and discordant) is approximately zero. Also, the simulated data examples illustrate that the proportion of differentially expressed genes, as estimated by the posterior mean of ξ, is typically calibrated. Nevertheless, in the experimental data example we observed a ξ of 0.77. Of these, 35% are predicted to be discordant and 42% are predicted to be concordant, and the distribution of these posterior probabilities show a large number of genes with values below .5 (Figure 6). Better calibration of ξ is a future direction of this research and may be possible by exploring different priors for the Δ.

Our Bayesian model can be modified and generalized in several respects. First, the possibility of missing gene expression observations can easily be included. The missing x_{gsp} can simply be integrated out from the posterior distribution. Second, in the current model we have assumed the same set of genes in all the studies. Partly overlapping gene sets can of course be included in the model just by considering expression values corresponding to genes that are not present in a study as missing. However, to design a more efficient computational algorithm one should integrate out both these x_{gsp}'s and the corresponding $\Delta'_{gp}s$ from the model, as is possible to do. Third, it is also technically straightforward to allow for missing observations in the phenotype. In that case we need to assign an additional probabilistic model for the ψ_{sp}'s and simulate the unobserved ones within the Metropolis–Hastings algorithm. This will in effect produce a prediction of the unobserved clinical variables. However, if the number of unobserved clinical variables is large, we expect it to be necessary to use block updates in the Metropolis–Hastings algorithm to avoid slow convergence and mixing.

Our results provide a strong indication that borrowing strength across both genes and studies can be effective in the analyis of multi-platform studies. As is the case for most complex multilevel

models, this comes at the price of added computational effort, and an increased burden of proof that the modeling assumptions are tenable.

References

Baldi, P. and Long, A. D. "A Bayesian framework for the analysis of microarray expression data: Regularized t–test and statistical inferences of gene changes." *Bioinformatics*, 17(6):509–519 (2001).

Barnard, J., McCulloch, R. R., and Meng, X.-L. "Modelling Covariance Matrices in Terms of Standard Deviations and Correlations, with Application To Shrinkage." *Statistica Sinica*, 10:1281–1311 (2000).

Beer, D. G., Kardia, S. L., Huang, C.-C., Giordano, T. J., andDavid E. Misek, A. M. L., Lin, L., Chen, G., and et al. "Gene-expression profiles predict survival of patients with lung adenocarcinoma." *Nature Medicine*, 8(8):816–824 (2002).

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J., and Meyerson, M. "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses." *Proceedings of the National Academy of Sciences USA*, 98:13790–13795 (2001).

Bröet, P., Richardson, S., and Radvanyi, F. "Bayesian hierarchical model for identifying changes in gene expression from microarray experiments." *Journal of Computational Biology*, 9:671–683 (2002).

Caffo, B., Dongmei, L., and Parmigiani, G. "Power conjugate multilevel models with applications to genomics." Technical report 62:2004, Johns Hopkins University, Dept. of Biostatistics (2004).

Choi, H., Shen, R., Chinnaiyan, A., and Ghosh, D. "A Latent Variable Approach for Meta-Analysis of Gene Expression Data from Multiple Microarray Experiments." *BMC Bioinformatics*, 8(1):364 (2007).
URL http://dx.doi.org/10.1186/1471-2105-8-364

Choi, J. K., Yu, U., Kim, S., and Yoo, O. J. "Combining multiple microarray studies and modeling interstudy variation." *Bioinformatics*, 19-1:I84–I90 (2003).

Conlon, E. "A Bayesian mixture model for metaanalysis of microarray studies." *Funct Integr Genomics* (2007).
URL http://dx.doi.org/10.1007/s10142-007-0058-3

Conlon, E., Song, J., and Liu, J. "Bayesian models for pooling microarray studies with multiple sources of replications." *BMC Bioinformatics*, 7(1):247 (2006).
URL http://dx.doi.org/10.1186/1471-2105-7-247

Conlon, E. M., Song, J. J., and Liu, A. "Bayesian meta-analysis models for microarray data: a comparative study." *BMC Bioinformatics*, 8:80 (2007).
URL http://dx.doi.org/10.1186/1471-2105-8-80

Consortium, M. A. Q. C., Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., de Longueville, F., Kawasaki, E. S., Lee, K. Y., Luo, Y., Sun, Y. A., Willey, J. C., Setterquist, R. A., Fischer, G. M., Tong, W., Dragan, Y. P., Dix, D. J., Frueh, F. W., Goodsaid, F. M., Herman, D., Jensen, R. V., Johnson, C. D., Lobenhofer, E. K., Puri, R. K., Schrf, U., Thierry-Mieg, J., Wang, C., Wilson, M., Wolber, P. K., Zhang, L., Amur, S., Bao, W., Barbacioru, C. C., Lucas, A. B., Bertholet, V., Boysen, C., Bromley, B., Brown, D., Brunner, A., Canales, R., Cao, X. M., Cebula, T. A., Chen, J. J., Cheng, J., Chu, T.-M., Chudin, E., Corson, J., Corton, J. C., Croner, L. J., Davies, C., Davison, T. S., Delenstarr, G., Deng, X., Dorris, D., Eklund, A. C., hui Fan, X., Fang, H., Fulmer-Smentek, S., Fuscoe, J. C., Gallagher, K., Ge, W., Guo, L., Guo, X., Hager, J., Haje, P. K., Han, J., Han, T., Harbottle, H. C., Harris, S. C., Hatchwell, E., Hauser, C. A., Hester, S., Hong, H., Hurban, P., Jackson, S. A., Ji, H., Knight, C. R., Kuo, W. P., LeClerc, J. E., Levy, S., Li, Q.-Z., Liu, C., Liu, Y., Lombardi, M. J., Ma, Y., Magnuson, S. R., Maqsodi, B., McDaniel, T., Mei, N., Myklebost, O., Ning, B., Novoradovskaya, N., Orr, M. S., Osborn, T. W., Papallo, A., Patterson, T. A., Perkins, R. G., Peters, E. H., Peterson, R., Philips, K. L., Pine, P. S., Pusztai, L., Qian, F., Ren, H., Rosen, M., Rosenzweig, B. A., Samaha, R. R., Schena, M., Schroth, G. P., Shchegrova, S., Smith, D. D., Staedtler, F., Su, Z., Sun, H., Szallasi, Z., Tezak, Z., Thierry-Mieg, D., Thompson, K. L., Tikhonova, I., Turpaz, Y., Vallanat, B., Van, C., Walker, S. J., Wang, S. J., Wang, Y., Wolfinger, R., Wong, A., Wu, J., Xiao, C., Xie, Q., Xu, J., Yang, W., Zhang, L., Zhong, S., Zong, Y., and Slikker, W. "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements." *Nat Biotechnol*, 24(9):1151–1161 (2006).
URL http://dx.doi.org/10.1038/nbt1239

Dellaportas, P. and Roberts, G. O. "An introduction to MCMC." In Møller, J. (ed.), *Spatial Statistics and Computational Methods*, number 173 in Lecture Notes in Statistics, 1–41. Springer, Berlin (2003).

Do, K.-A., Müller, P., and Tang, F. "A Bayesian mixture model for differential gene expression." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):627–644 (2005).
URL http://www.blackwell-synergy.com/doi/abs/10.1111/j.1467-9876.2005.05593.x

Dongmei Liu, G. P. and Caffo, B. "Screening for Differentially Expressed Genes: Are Multilevel Models Helpful?" Technical Report 34, Johns Hopkins University Department of Biostatistics Working Paper (2004).

Efron, B. and Tibshirani, R. "Empirical Bayes methods and false discovery rates for microarrays." *Genetic Epidemiology*, 23(1):70–86 (2002).

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. "Empirical Bayes Analysis of a Microarray Experiment." *Journal of the American Statistical Association*, 96:1151–1160 (2001).

Farmer, P., Bonnefoi, H., Becette, V., Tubiana-Hulin, M., Fumoleau, P., Larsimont, D., Macgrogan, G., Bergh, J., Cameron, D., Goldstein, D., Duss, S., Nicoulaz, A.-L., Brisken, C., Fiche, M., Delorenzi, M., and Iggo, R. "Identification of molecular apocrine breast tumours by microarray analysis." *Oncogene*, 24(29):4660–4671 (2005).
URL http://dx.doi.org/10.1038/sj.onc.1208561

Garber, M. E., Troyanskaya, O. G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., van de Rijn, M., Rosen, G. D., Perou, C. M., Whyte, R. I., Altman, R. B., Brown, P. O., Botstein, D., and Petersen, I. "Diversity of gene expression in adenocarcinoma of the lung." *Proceedings of the National Academy of Sciences USA*, 98:13784–13789 (2001).

Garrett-Mayer, E., Parmigiani, G., Zhong, X., Cope, L., and Gabrielson, E. "Cross-study validation and combined analysis of gene expression microarray data." *Biostatistics* (2007).
URL http://dx.doi.org/10.1093/biostatistics/kxm033

Garrett-Mayer, E., Parmigiani, G., Zhoung, X., Cope, L., and Gabrielson, E. "Cross-study validation and combined analysis of gene expression microarry data." Technical report 65:2004, Johns Hopkins University, Dept. of Biostatistics (2004).

Gentleman, R., Ruschhaupt, M., and Huber, W. "On the synthesis of microarray experiments." Technical Report 8, The Berkeley Electronic Press, http://www.bepress.com/bioconductor/paper8 (2005).

Ghosh, D., Barette, T. R., Rhodes, D., and Chinnaiyan, A. M. "Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer." *Funct Integr Genomics*, 3(4):180–8 (2003). 1438-793X (Print) Journal Article Meta-Analysis.

Gottardo, R., Pannucci, J., Kuske, C., and Brettin, T. "Statistical analysis of microarray data: a Bayesian approach." *Biostatistics*, 4:597–620 (2003).

Green, P. "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82:711–732 (1995).

Hastings, W. "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika*, 57:97–109 (1970).

Hayes, D. N., Monti, S., Parmigiani, G., Gilks, C. B., Naoki, K., Bhattacharjee, A., Socinski, M. A., Perou, C., and Meyerson, M. "Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts." *J Clin Oncol*, 24(31):5079–5090 (2006).
URL http://dx.doi.org/10.1200/JCO.2005.05.1748

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O., Wilfond, B., Borg, A., and Trent, J. "Gene-expression profiles in hereditary breast cancer." *N Engl J Med*, 344(8):539–48 (2001).

Hedges, L. V. and Olkin, I. *Statistical Methods for Meta-analysis*. Academic Press (1985).

Huang, E., Cheng, S. H., Dressman, H., Pittman, J., Tsou, M. H., Horng, C. F., Bild, A., Iversen, E. S., Liao, M., Chen, C. M., West, M., Nevins, J. R., and Huang, A. T. "Gene expression predictors of breast cancer outcomes." *Lancet*, 361(9369):1590–6 (2003).

Ibrahim, J. G., Chen, M. H., and Gray, R. J. "Bayesian Models for Gene Expression with DNA Microarray Data." *Journal of the American Statistical Association*, 97:88–99 (2002).

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." *Biostatistics*, 4(2):249–264 (2003).
URL http://dx.doi.org/10.1093/biostatistics/4.2.249

Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., Gabrielson, E., Garcia, J. G. N., Geoghegan, J., Germino, G., Griffin, C., Hilmer, S. C., Hoffman, E., Jedlicka, A. E., Kawasaki, E., Martnez-Murillo, F., Morsberger, L., Lee, H., Petersen, D., Quackenbush,

J., Scott, A., Wilson, M., Yang, Y., Ye, S. Q., and Yu, W. "Multiple-laboratory comparison of microarray platforms." *Nat Methods*, 2(5):345–350 (2005).
URL http://dx.doi.org/10.1038/nmeth756

Ishwaran, H. and Rao, J. "Detecting differentially expressed genes in microarrays using Bayesian model selection." *Journal of the American Statistical Association*, 98(462):438–455 (2003).

—. "Spike and slab gene selection for multigroup microarray data." *Journal of the American Statistical Association*, 100(471):764–780 (2005).

Johnson, W., Li, C., and Rabinovic, A. "Adjusting batch effects in microarray data using empirical Bayes methods." *Biostatistics*, 8(1):118–127 (2007).

Jung, Y.-Y., Oh, M.-S., Shin, D. W., Kang, S.-H., and Oh, H. S. "Identifying differentially expressed genes in meta-analysis via Bayesian model-based clustering." *Biom J*, 48(3):435–450 (2006).

Kendziorski, C., Newton, M., Lan, H., and Gould, M. "On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles." *Statistics in Medicine*, 22:3899–3914 (2003).

Kerr, K. "Extended analysis of benchmark datasets for Agilent two-color microarrays." *BMC Bioinformatics*, 8(1):371 (2007).
URL http://dx.doi.org/10.1186/1471-2105-8-371

Liu, D., Parmigiani, G., and Caffo, B. "Screening for differentially expressed genes: Are multilevel models helpful?" Technical report 34:2004, Johns Hopkins University, Dept. of Biostatistics (2004).

Lönnstedt, I. and Speed, T. "Replicated Microarray Data." *Statistica Sinica*, 12(1):31–46 (2002).

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. "Equations of state calculations by fast computing machine." *J. Chem. Phys.*, 21:1087–1091 (1953).

Modrek, B. and Lee, C. "A genomic view of alternative splicing." *Nat Genet*, 30(1):13–19 (2002).
URL http://dx.doi.org/10.1038/ng0102-13

Newton, M., Noueiry, A., Sarkar, D., and Ahlwuist, P. "Detecting differential gene expression with a semiparametric hierarchical mixture method." *Biostatistics*, 5:155–176 (2004).

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. "On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data." *Journal of Computational Biology*, 8:37–52 (2001).

Pan, W. "A Comparative Review of Statistical Methods for Discovering Differentially Expressed Genes in Replicated Microarray Experiments." *Bioinformatics*, 18:546–554 (2002).

Parmigiani, G. "Measuring uncertainty in complex decision analysis models." *Stat Methods Med Res*, 11(6):513–537 (2002).

Parmigiani, G., Garrett-Mayer, E. S., Anbazhagan, R., and Gabrielson, E. "A cross-study comparison of gene expression studies for the molecular classification of lung cancer." *Clin Cancer Res*, 10(9):2922–2927 (2004).

Rhodes, D. R., Barrette, T. R., Rubin, M. A., Ghosh, D., and Chinnaiyan, A. M. "Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer." *Cancer Res*, 62(15):4427–33 (2002). 0008-5472 (Print) Journal Article Meta-Analysis.

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., and Chinnaiyan, A. M. "Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression." *Proc Natl Acad Sci U S A*, 101(25):9309–14 (2004). 0027-8424 (Print) Journal Article Meta-Analysis.

Shen, R., Ghosh, D., and Chinnaiyan, A. "Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data." *BMC Genomics*, 5(1):94 (2004).
URL http://www.biomedcentral.com/1471-2164/5/94

Smith, A. F. M. and Roberts, G. O. "Bayesian Computation Via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods (Disc: P53-102)." *Journal of the Royal Statistical Society, Series B, Methodological*, 55:3–23 (1993).

Smyth, G. K. and Speed, T. "Normalization of cDNA microarray data." *Methods*, 31(4):265–273 (2003).

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Eystein Lonning, P., and Borresen-Dale, A. L. "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications." *Proc Natl Acad Sci U S A*, 98(19):10869–74 (2001).

Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X.-W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., Lee, C., Montie, J. E., Shah, R. B., Pienta, K. J., Rubin, M. A., and Chinnaiyan, A. M. "Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer." *Science*, 310(5748):644–648 (2005).
URL http://dx.doi.org/10.1126/science.1117679

Townsend, J. and Hartl, D. "Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple treatements or samples." *Genome Biology*, 3:research0071.1–71.16 (2002).

Tseng, G., Oh, M., Rohlin, L., Liao, J., and Wong, W. "Issues in cDNA Microarray Analysis: Quality Filtering, Channel Normalization, Models of Variations and Assessment of Gene Effects." (2001). Submitted to Nucleic Acids Research.

Tusher, V., Tibshirani, R., and Chu, G. "Significance analysis of microarrays applied to the ionizing radiation response." *Proceedings of the National Academy of Sciences*, 98:5116–5121 (2001).

Wang, J., Coombes, K. R., Highsmith, W. E., Keating, M. J., and Abruzzo, L. V. "Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies." *Bioinformatics*, 20(17):3166–3178 (2004).
URL http://dx.doi.org/10.1093/bioinformatics/bth381

Wu, L. F., Hughes, T. R., Davierwala, A. P., Robinson, M. D., Stoughton, R., and Altschuler, S. J. "Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters." *Nat Genet*, 31(3):255–265 (2002).
URL http://dx.doi.org/10.1038/ng906

Zhong, X., Marchionni, L., Cope, L., Iversen, E. S., Garrett-Mayer, E. S., Gabrielson, E., and Parmigiani, G. "Optimized Cross-study analysis of microarray based predictors." Technical Report 129, johns Hopkins University Department of Biostatistics (2007).
URL http://www.bepress.com/jhubiostat/paper129/

9 Supplemental Material

	hyperparameters
α_a	1
β_a	1
p_a^0	0.1
p_a^1	0.1
α_b	1
β_b	1
p_b^0	0.1
p_b^1	0.1
ν_r	P + 1
ν_ρ	P + 1
α_ξ	1
β_ξ	1
$c^2_{\max}$	1

Table 4: Default values of the hyperparmeters in the Bayesian model for a dataset with P studies. For each of the simulated and real datasets in this paper, we used the default hyperparameters with $P = 3$.

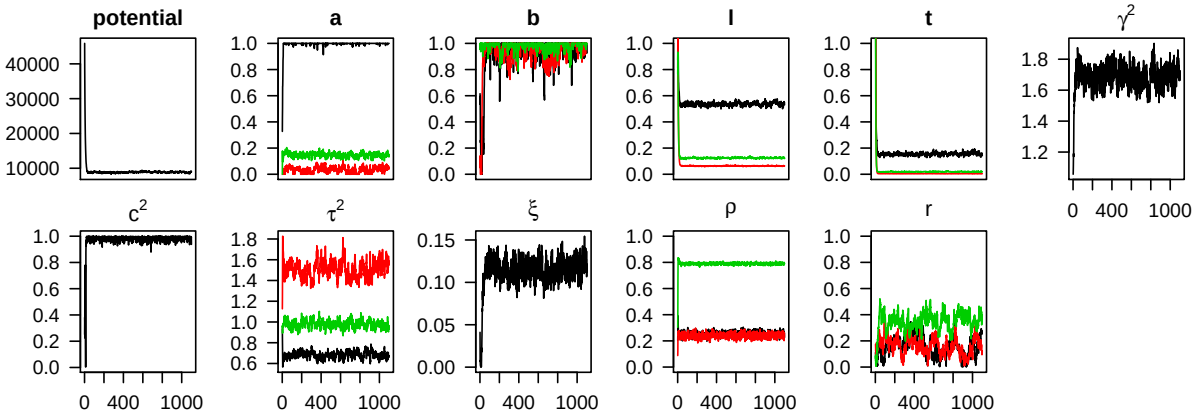
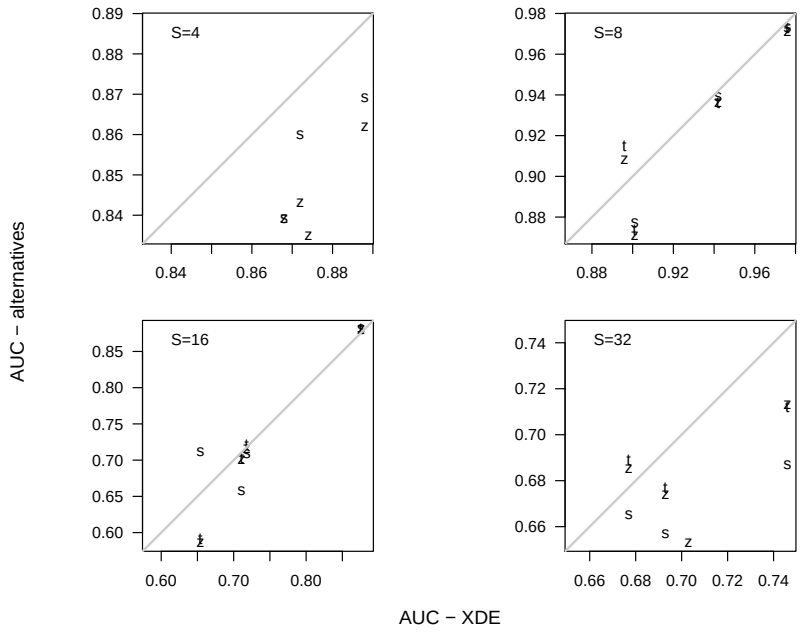
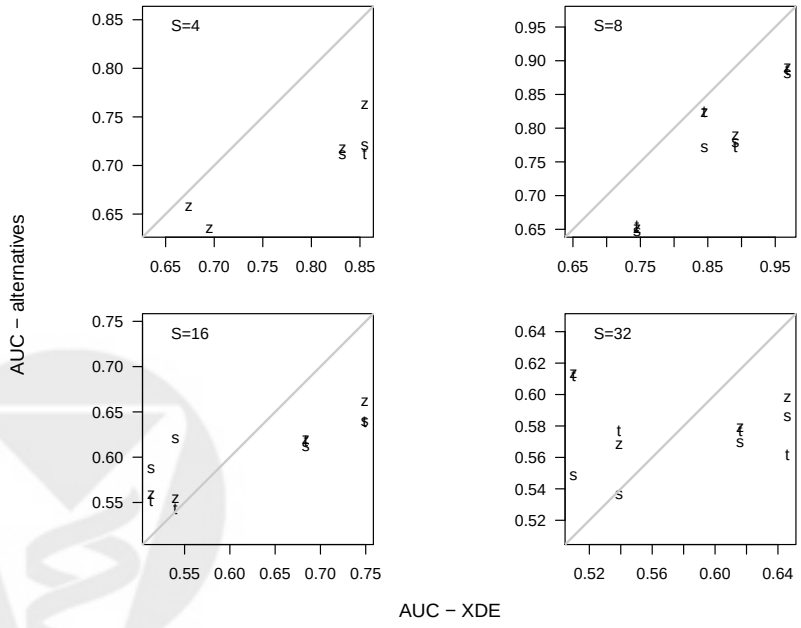


Figure 9: Trace plots of Metropolis-Hastings parameters obtained from fitting the Bayesian model to a simulated dataset of three studies. The parameters used to simulate the data are provided in row 1 of Table 1. A thinning interval of 20 was used in this plot, hence only 1100 out of 22,000 iterations are plotted. The first 4000 iterations were discarded before calculating posterior statisics of interest.



(a) Differential expression ($\mathcal{E}$)



(b) Discordant differential expression ($\mathcal{D}$)

Figure 10: In each panel, we plot the AUC obtained from alternative methods on the vertical axis and the AUC from the Bayesian model (*XDE*) on the horizontal axis. Differential expression ($\mathcal{E}$) and discordant differential expression ($\mathcal{D}$) are considered separately. In several instances, the AUC corresponding to the `t`- and `SAM`-scores were lower than the limit used for the scatterplots and were not plotted.

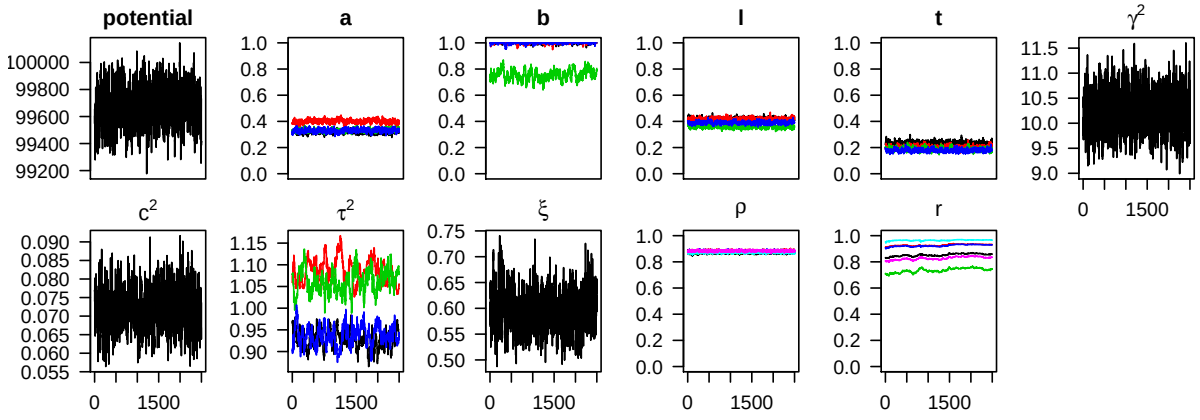


Figure 11: A single dataset, the `Huang` study, was split into four disjoint parts with 5 ER negative and 16 ER positive samples in each. Plotted are traces for a subset of the Metropolis-Hastings parameters.

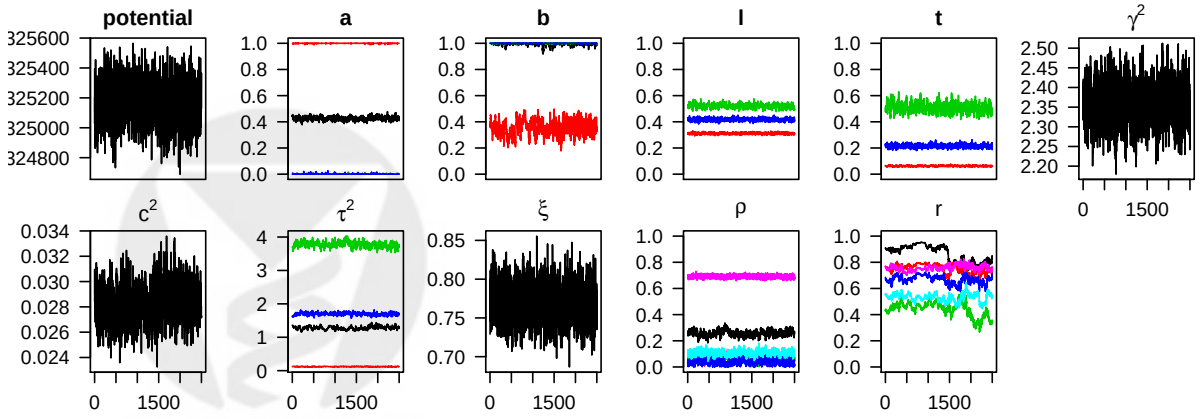


Figure 12: Traceplots for Metropolis-Hastings parameters in the experimental data example.